INRIA

CENTRE DE ROCQUENCOURT

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Voluceau
Rocquencourt
B.P.105
78153 Le Chesnay Cedex
France

Tél. : (1) 39 63 55 11

Rapports de Recherche

N° 502

# LIMITING DISTRIBUTIONS FOR TRIE PARAMETERS

Philippe JACQUET
Mireille REGNIER

Mars 1986

# LIMITING DISTRIBUTIONS FOR TRIE PARAMETERS

*Philippe JACQUET*

*Mireille REGNIER*

INRIA
Rocquencourt
78153-Le Chesnay (France)

ABSTRACT

*We consider here the well-known trie structure. We consider three basic parameters: the size, the depth of the leaves and the height of a trie that is formed with n data. Our first result is the proof of the convergence of their distributions and of their moments of any order when $n \to \infty$. We exhibit the limits: a periodic distribution or a normal distribution. The results are given for uniform or biased data distributions for Bernoulli and Poisson models. Our reasoning is based on generating and characteristic functions. We make a great use of analytic functions and asymptotic methods. Our second result is a tauberian theorem, a quite general tool to translate results under the Poisson model to the Bernoulli model.*

**Keywords:** tries, algorithms, distributions, generating functions.

This paper provides a uniform framework to establish limiting distributions for some of the more important trie parameters. Tries are a basic tree structure associated to a recursive partitioning process which appears in quite a large number computing problems. A trie may be used as an index to access data on secondary memory. This is Dynamic Hashing Algorithms[13, 12, 2] , or k-d-tries and Grid-File Algorithms in the multidimensional case [15] . Tries also are an underlying structure in problems as various as: communication protocols, polynomial factorization, radix exchange sort, simulation algorithms, Huffman's algorithm...[3, 9, 11] . Usual tree parameters such as the **size, external path length** or **height** have in these applications a simple interpretation as a cost of execution as it will be detailed below in Section 1.

Our purpose here is twofold. First, we prove the ***convergence to a simple limiting distribution*** (gaussian or periodic) of the distribution of basic trie parameters like the size, the external path length or the height. In passing, we get the asymptotics for the moments of any order, and in particular, the mean and the variance. This extends some previous results. [5, 1, 6, 10, 17] . In these papers, exact expressions of some average values and, occasionally, of variances are given and asymptotic developments are derived. We also solve a conjecture in[18] . We assume natural hypotheses on the number and the distribution of the data and use the so-called *Bernoulli* model -uniform or

RESUME :

Nous examinons la structure usuelle arborescente dite de "trie". Nous considérons trois paramètres fondamentaux : la taille, la profondeur des feuilles et la hauteur quand le trie est développé à partir de n insertions. Sous l'hypothèse que la distribution des clefs correspondant aux n insertions obeissent aux statistiques naturelles de Bernoulli ou de Poisson, nous montrons la convergence des distributions de ces paramètres vers des lois simples quand n croît indéfiniment. Nous basons notre analyse sur les fonctions génératrices et caractéristiques des distributions considérées. Nous utilisons abondamment l'analyse complexe pour nos évaluations asymptotiques. Chemin faisant nous établissons un théorème général "semi tauberien" qui s'avère être un outil précieux pour effectuer la liaison entre les modèles de Bernoulli et de Poisson.

biased- and its approximation, the *Poisson* model (defined below). Second, we prove a ***semi-tauberian*** theorem that allows to deduce from the asymptotics of a generating function the asymptotics of its coefficients. This theorem is of great use here as, applying it, the Poisson convergence translates easily to the Bernoulli model.

In our proofs, we shall first study the Poisson model which is technically easier to deal with than the Bernoulli model, although less accurate. Our basic tool is the **generating functions** and the **characteristic functions.** The convergence proof is amenable to the study of the analytical properties and the growth of the generating functions. A recurrence functional equation is derived that one cannot solve explicitly. Thus, we study the asymptotic properties of the generating functions involved. To do so, we eventually reduce a non linear equation to a ***quasi-linear*** one. Then, we use the ***Mellin Transform***[8, 7] in order to get the asymptotic properties and the limiting form of the characteristic function when the size $n$ of the data structure tends to $\infty$. Then, by the Continuity Theorem, we get the convergence of the distribution in the Poisson model. Finally, using the tauberian theorem proved in this paper, we derive the limiting distribution in the Bernoulli case. This ***Tauberian-like*** reasoning is based on integral calculations. In particular, we introduce a new integral transform $Z$ and prove an extended version of the Stationary Phase Theorem.

The plan of the paper is the following. In the first Section, we briefly define the trie structure and the Bernoulli and Poisson statistical assumptions under which they are analysed. We also give an algorithmic interpretation of the parameters to be studied. In Section 2, we present a general sketch of the derivation in the Poisson case. We treat the distribution of the size of the tries, which provides a good example of a ***quasi-linearisation.*** Section 3 is devoted to the derivation of a general tauberian theorem. We show how it applies to the size of the tries. In Section 4, we deal with the ***depth of the leaves*** and the ***internal path length.*** This case is of particular interest as the nature of the limiting law -periodic or gaussian- is different whether the data distribution is biased or uniform. In Section 5, we consider the ***height*** of the tries, using an ***approximation*** of the generating function. Section 6 is a conclusion.

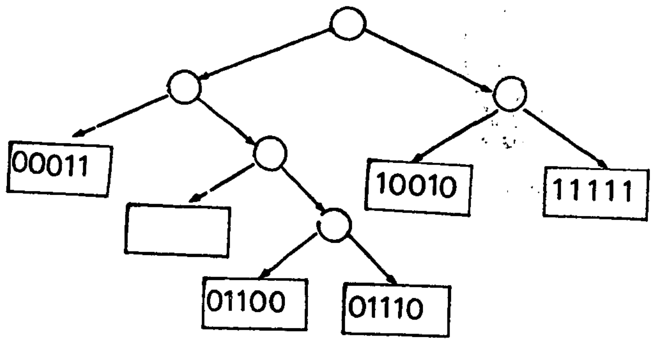
## I. DEFINITIONS.

### I.1. DEFINITION AND APPLICATIONS OF THE TRIE STRUCTURE:

Tries are a common data structure which allows the representation as a tree of digital (e.g. binary) sequences [11] . A trie can be naturally associated to a set $S$ of binary sequences by a recursive partitioning into subsets of smaller cardinality at most one: whenever $|S|\geq 2$, a root is created; left and right subtrees are then associated to subsets $S_0$ and $S_1$ consisting of the keys of $S$ whose first bits are respectively 0 and 1. The process is repeated on $S_0$ (and $S_1$), using the second bit of sequences,etc... So, generally, at the $i$-th step, we use the $i$-th bit. The process is stopped when one finds subsets of cardinality at most 1 which are represented as leaves. For example, to the set:

$$S = \{00011, 01100, 01110, 10010, 11111\}$$

is associated the trie:




Notice that the same algorithm may be used to individuate subsets of cardinality at most $b$, by stopping the recursive partitioning when sets of cardinality at most $b$ have been found. Moreover, the trie adapts well to dynamically growing sets of data. Operations like queries, insertions or deletions of a key sequence can be performed efficiently (in expected time close to $\log_2 n$). To retrieve an element, one follows in the trie a path to a leaf determined by the successive bits of the element to be found. Checking the ***presence*** of the key in the leaf, one may answer a query. One may also insert a key: if the leaf is not full, use it, else, split the subset of keys found there, according to the general rule.

The costs of operations performed on a trie, for different applications, can be expressed as functions of the classical parameters defined on trees. We consider in the following the distribution of three parameters. Namely, **the size, the depth of the leaves and the height** of a random trie.

Now, we make precise our statistical assumptions. [4] .

We assume that the keys are ***infinite sequences*** of bits. Thus they belong to $\{0,1\}^\infty$ or, equivalently, range in the real interval $[0,1]$. We also assume that the bit sequences are sequences of independent Bernoulli trials with the same probability $p$. *I.e.* the probability for the $i$-th bit to be 0 or 1 is:

$$P(b_i=0)=p,\ P(b_i=1)=q=1-p.$$

When $p=q=\frac{1}{2}$, the model is ***uniform***: keys are uniformly drawn from $[0,1]$. When $p \neq q$, the model is ***biased*** with bias $p$. Now, in the ***Bernoulli model***, the cardinality of the data set is a ***fixed*** number $n$. A good approximation is provided by the ***Poisson model***. The cardinality of the data set is then a ***random variable*** $N$ following a Poisson law with parameter $n$ (For reasons to be explained later, we shall alternatively use the parameter $z$ to denote the parameter of the Poisson law). Equivalently, the probability that the structure is formed with $m$ records is:

Notice that such a distribution is concentrated around its mean $n$ (or $z$).

We are now ready to give an algorithmic interpretation of our three parameters for the main computing problems involving tries.

The **communication protocol** of Capetanakis-Tsybachov-Mikhailov assumes that several transmitters are sharing a unique channel to send information. If at some instant $n$ users, $n>1$ , try to transmit, they "collide". The packets are destroyed and the transmission fails. In such a case, the algorithm suggests users should start playing independently head and tail and thus be associated different strings 0−1. A trie may be built and prefix order determines the order users are allowed to transmit. The algorithm works whether the dices are loaded -all the same- or not. There are slightly modified algorithms which are more efficient when $p \neq q$. The size is exactly the *length of a session* that separates $n$ users. It is also the time the channel is closed to others transmitters. The depth of the leaves is a tool to analyze the *waiting time* of customers.

The trie structure also appears in **file addressing on secondary memory**, for example in databases. Datas are characterized by one or several *keys*. These keys -or their binary representation- can be assumed to be infinite sequences from $\{0,1\}^*$. In this case, $b$ stands for the capacity of the pages in secondary memory and a trie is used as an index: in the leaves one finds pointers to the data. Then, retrieval amounts to a traversal of the trie. This structure supports dynamic modifications, as insertions and deletions only modify slightly the partition and the trie. When a single key is used, we get exactly the partition process described above, see *Dynamic* and *Extendible Hashing*. When several keys are used, shuffling them reduces to that scheme: see *Grid-File Algorithms*. In monodimensional and multidimensional cases, we can assume biased or uniform distributions. A nice generalization of the biased case is provided when keys are strings of characters[14] . Datas may be characterized by a set of keywords whose concatenation makes a key. Then, using some binary representation, one can build a trie as an index. Notice that the probabilities $P(a,b)$ for finding a $b$ after an $a$ in a word are not equal and they induce biased probabilities that depend on the level. We have then a *markovian process*. The results we will establish when $p$ and $q$ are independent of the level apply here as well. The size of the trie stands for the *space* occupied in secondary memory. The height provides an indication on the space necessary to store the index. The depth of insertion allows for *cost evaluation* of some operations performed in central memory, as the traversal of the index.

For **radix exchange sort**, the depth of insertion stands for the number of bits comparisons to be done during an insertion.

## 1.2. NOTATIONS AND BASIC THEOREMS:

We define here generating functions and state a relationship between the distributions of random variables under the Bernoulli and Poisson models. We also recall basic theorems on the convergence of distributions.

**Definition:**

*Let $Y$ be a discrete parameter. Let $P_n^k$ be the probability for $Y$ to be equal to $k$ when the Bernoulli model is used and the size of the data set is $n$. We note:*

$$\begin{cases} P_n(u) = \sum_k P_n^k u^k \\ P(z,u) = e^{-z} \sum_n \frac{z^n}{n!} P_n(u) \end{cases}$$

*The first formal series is the ordinary generating function. The second is called the Poisson generating function.*

**Remark** *A priori*, these expressions are well defined for $|u|<1$ and any $z$ in $C$. When we set $u=e^{it}, t \in R, P_n(e^{it})$ is the **characteristic function** of the distribution under the Bernoulli model with first parameter $n$. When $z$ is real and positive, $P(z,e^{it})$ is the characteristic function of the parameter $Y$ when the number of records in the trie follows a Poisson law with mean $z$. As a matter of fact, the Poisson probability that a parameter $Y$ be $k$ is:

$$\sum_n P_n^k \frac{z^n}{n!} e^{-z} \ ;$$

thus the associated generating function is:

$$\sum_{n,k} P_n^k u^k \frac{z^n}{n!} e^{-z} = P(z,u)$$

Finally, one can express easily the moments of any order as a function of $P(z,u)$ and its derivatives. In particular, we set the definition:

**Definition:** Let $X(z)$ and $v(z)$ be the two formal series of the variable $z$:

$$X(z) = \frac{\partial P}{\partial u}(z,1)$$

$$v(z) = \frac{\partial^2 P}{\partial u^2}(z,1) + \frac{\partial P}{\partial u}(z,1) - [\frac{\partial P}{\partial u}(z,1)]^2.$$

**Remark:** For $z$ real, $X(z)$ and $v(z)$ are the mean and the variance of the parameter associated to $P$ under the Poisson model. We also note

$$\sigma(z) = v(z)^{\frac{1}{2}}$$

when $z \in R^+$.

To prove the convergence of the distributions and of their moments, we make use of the two following theorems :

**Theorem 1 (Continuity Theorem)**[4

*In order that a sequence $\{F_n\}$ of probability distributions converge properly to a probability distribution $F$ it is necessary and sufficient that the sequence $\{\varphi_n\}$ of their characteristic functions converges pointwise to a limit $\varphi$, and that $\varphi$ is continuous at the origin.*

**Theorem 2:** *If a probability distribution $\varphi(t)$ is analytic in a neighbourhood of $R$, then it is fully determined by its derivatives in* 0 *or, equivalently, by its moments.*

Let us recall the definition:

**Definition:** *The normal -or gaussian- distribution is the distribution associated to the continuous characteristic function:*

$$\varphi(t) = e^{\frac{-t^2}{2}}$$

Thus, to get the conditions of Theorem 2, we shall consider $t$ as a complex variable. We shall also consider $z$ as a complex variable. Then, via a tauberian theorem, we will be able to derive Bernoulli results from Poisson results.

## II. LIMITING DISTRIBUTIONS UNDER THE POISSON MODEL:

We are dealing here with the Poisson model and prove the convergence to the normal distribution of the distribution of the size of the tries, i.e. the number of internal nodes. This provides a general sketch for the proofs of the paper. Throughout this section, $P$ will be the Poisson generating function associated to that parameter. At the first step, we derive a functional equation satisfied by $P$. Then, we prove the analyticity of $P(z,u)$ with respect to $z$ and $u$, and study its analytical properties when $z \to \infty$. Finally, to modify the equation,we consider $\log P(z,u)$ and prove some analytical and growth properties. At the second step, we study the asymptotic properties of $X(z)$ and $v(z)$. Combining this with the previous results, we establish the convergence to the normal distribution.

We said previously that $z$ and $u$ were complex variables. It will prove convenient to have $z$ varying in cones with vertex 0. Thus, we note:

$$C_{A,\vartheta} = \{z\,;\, A<Re(z)\ ,\ |Arg(z)|<\vartheta\}.$$

All over this paper, the notation $V(\alpha)$ (or, eventually, $V_\alpha(\alpha)$ ) represents a neighbourhood of $\alpha$.

We prove first Proposition II.1.:

**Proposition II.1.:** *Let $P(z,u)$ be the generating function for the size of random tries. $P(z,u)$ is defined and analytical with respect to $z$ and $u$, in the domain* $\mathbf{C}\times B(0,\frac{1}{p^{b+1}+q^{b+1}})$, *where* $B(0,\frac{1}{p^{b+1}+q^{b+1}})$ *is the open ball with center 0 and radius* $\frac{1}{p^{b+1}+q^{b+1}}$. *It satisfies there the functional equation:*

$$P(z,u) = uP(pz,u)P(qz,u) + (1-u)e_b(z)e^{-z} \tag{E1}$$

*and has the following growth properties:*

(P1) *For any compact set $(D)$ in* $B(0,\frac{1}{p^{b+1}+q^{b+1}})$, *there exists a constant $\beta$ such that:*

$$|P_n(u)|<\beta^{n-1},$$

(P2) $\forall\alpha:0<\alpha<1, \forall\vartheta\in[0,\frac{\pi}{2}[, \forall c:0<c<1$ *there exist a neighbourhood $V_\alpha(1)$ and a cone $C_{A,\vartheta}$ such that:*

$$z\in C_{A,\vartheta}\ ,\ u\in V_\alpha(1) \Rightarrow |P(z,u)|>c\,|e^{-\alpha z}|.$$

**Proof:** According to the recursive definition of tries, the number $I$ of internal nodes of a trie $t$ with left subtrie $t_0$ and right subtrie $t_1$ satisfies:

$$I(t) = I(t_0) + I(t_1) + 1-\chi_{|t|\le b}.$$

Here, $\chi(P)$ is the characteristic function of property $P$ (i.e. $\chi(P)=1$ if $P$ is true, $\chi(P)=0$ otherwise). Using the algebraic methods defined in [6] , one may derive "directly" $(E1)$. From $(E1)$, one gets the recurrence:

$$P_0(u) = \cdots = P_b(u) = 1$$

$$P_n(u)\,(1-u(p^n+q^n)) = u\sum_{0<j<n}\binom{n}{j}p^jP_j(u)q^{n-j}P_{n-j}(u),$$

which shows that the $P_n(u)$ are rational fractions and insures the analyticity of $P_n(u)$ in $B(0,\frac{1}{p^{b+1}+q^{b+1}})$. Moreover, noting $\beta = \max_D(\frac{|u|}{|1-u(p^{b+1}+q^{b+1})|},1)$, we get $(P1)$ by induction. The analyticity of $P$ follows.

Now we prove $(P2)$ by induction. To start the recurrence, we choose $A$ such that:

$$e^{\alpha A\cos\vartheta}>\frac{2}{c}, \text{ and } \sup_{C_{A,\vartheta}}|e_b(z)e^{-(1-\alpha)z}|<1.$$

Now, since $P(z,1)$ is 1 and $P$ is uniformly continuous on compact sets, there exists $V_\alpha(1)$ such that, in the domain $D=\{z\in C_{A,\vartheta}, Re(z)\le\frac{A}{p}\}\times V_\alpha(1)$, $P$ satisfies:

$$|P(z,u)|\ge c > c\,|e^{-\alpha z}|.$$

Now, assuming $p<q$, let $D_m$ be the domains $\{z\in C_{A,\vartheta}, A\le Re(z)\le\frac{A}{pq^m}\}\times V_\alpha(1)$. They satisfy:

$$(z,u)\in D_{m+1}-D_0 \Rightarrow \{(pz,u)\in D_m,(qz,u)\in D_m\}.$$

For $(z,u)$ in $D_0$, $|P(z,u)e^{\alpha z}|>2$ holds. Assuming $|u-1|<1$ and $|u|>\frac{3}{4}$, one gets, for (z,u) in $D_{m+1}-D_0$:

$$|P(z,u)|>\frac{3}{4}.4|e^{-\alpha(p+q)z}| - |e^{-\alpha z}| = 2|e^{-\alpha z}| > c\,|e^{-\alpha z}|.$$

■

The equation $(E1)$ is rather intricate and does not provide directly asymptotics for $P(z,e^{it})$. Thus, we reduce it to an almost linear equation by considering $L(z,u)=\log P(z,u)$.

**Proposition II.2.:** *For any cone $C_{0,\vartheta}$, $0<\vartheta<\frac{\pi}{2}$, there exists a neighbourhood of $1, V(1)$, such that $L(z,u)=\log P(z,u)$ be defined and analytical when $(z,u)\in C_{0,\vartheta}\times V(1)$. It satisfies there:*

$$|L(z,u)|<A.|z|,$$

*uniformly in $u$.*

**Proof:** The existence and analyticity follow from Proposition II.1., as $P$ is analytic and non zero. Note:

$$g(z,u)=\log(1+\frac{(1-u)e_b(z)e^{-z}}{P(z,u)}), \quad \varphi(z,u)=\frac{L(z,u)-\log(u)}{z}.$$

Using $(E1)$ and $(P2)$, we get a "linear" form for $(E1)$:

$$\varphi(z,u)=p\,\varphi(pz,u)+q\,\varphi(qz,u)+\frac{g(z,u)}{z}. \qquad \text{(E2)}$$

with $g(z,u)<B.e^{\alpha z}e^{-z}$. Let now :

$$D_n=\{(z,u); u\in V(1), z\in C_{0,\vartheta}, Re(z)<\frac{1}{q^n}\}$$

be an increasing sequence of truncated cones, and $A_n$ be:

$$A_n=\sup_{D_n}|\varphi(z,u)|.$$

As the maximum on $D_n$ is obtained either for $z$ in $D_{n-1}$ or for z satisfying the equation above, with $pz$ and $qz$ in $D_{n-1}$, we get:

$A_n \leq A_{n-1} + B\, e^{-(1-\alpha)q^{-n}} \leq A_0 + B \sum_n e^{-(1-\alpha)q^{-n}} \leq A.$

■

We can now establish asymptotics for $P(z, e^{it})$ :

**Theorem II.3.:** *For any $C_{0,\vartheta}$, $0<\vartheta<\frac{\pi}{2}$, there exists a compact neighbourhood $V(0)$ of 0 in $C$ such that :*

$$P(z, e^{it}) = e^{itX(z) - \frac{v(z)}{2}t^2 + O(zt^3)}, (z,t) \in C_{0,\vartheta} \times V(0),$$

*$X(z)$ and $v(z)$ being $O(z)$ functions.*

**Proof:** $L$ is analytic with respect to $u = e^{it}$ and thus with respect to $t$. Computing its first derivatives, we get,for $u$ in $V(1)$ or $t$ in $V(0)$:

$$L(z,u) = iX(z)t - \frac{v(z)}{2}t^2 + g_z(t).t^3,$$

where $g_z(t)$ is analytic. Applying the Cauchy formulae, for every integer $p$, on a contour $C$ included in $V(0)$ and encircling 0:

$$f(t) = \frac{1}{2i\pi}\int_C \frac{f(\omega)}{\omega - t}dt$$

$$\frac{f^{(p)}(0)}{p!} = \frac{1}{2i\pi}\int_C \frac{f(\omega)}{\omega^{p+1}}d\omega$$

we get:

$$X(z) = \frac{1}{2i\pi}\int_C \frac{L(z,e^{i\omega})}{\omega^2}d\omega$$

$$\frac{v(z)}{2} = \frac{1}{2i\pi}\int_C \frac{L(z,e^{i\omega})}{\omega^3}d\omega$$

$$g_z(t) = \frac{1}{2i\pi}\int_C \frac{L(z,e^{i\omega})}{\omega^3(\omega - t)}d\omega,$$

where $\frac{1}{\omega - t}$ is upper bounded on the contour. Thus $X(z)$, $v(z)$ are $O(z)$ functions. Moreover, if $V(0)$ is redefined as a closed neighbourhood interior to $C$, $g_z(t)$ is also uniformly $O(z)$ in $t$.

To prove the convergence to the normal distribution, we need more precisions on the growth of $X(z)$ and $v(z)$.

**Lemma II.4.:** *The mean and variance $X(z)$ and $v(z)$ under the Poisson model satisfy the equations:*

$$\begin{cases} X(z) = X(pz) + X(qz) + 1 - e_b(z)e^{-z} \\ v(z) = v(pz) + v(qz) + (2X(z) - 1 + e_b(z)e^{-z})e_b(z)e^{-z}. \end{cases}$$

*Asymptotically, we have:*

$$\begin{cases} X(z) \sim z \; Q_1(z) \\ v(z) \sim z \; Q_2(z) \end{cases}$$

*where $Q_1$ and $Q_2$ are bounded by strictly positive constants. When $p=q=\frac{1}{2}$, $Q_1$ and $Q_2$ are also periodic in $\{\log_2(z)\}$.*

**Proof:** The equations are derived from the definitions and $(P1)$. The asymptotic values are derived via Mellin transform $X^*(s)$ and $v^*(s)$ of respectively $X(z)$ and $v(z)$,

$$X^*(s) = \frac{\Gamma(s+b)}{\Gamma(b)(1-p^{-s}-q^{-s})}$$

$$v^*(s) = \frac{M[(2X(z)-1+e_b(z)e^{-z})e_b(z)e^{-z};s]}{1-p^{-s}-q^{-s}}$$

by the methods developed in[8, 17] . More extended results are given in[8, 11, 3, 17] .

In Section 3, we will also need the related results:

**Lemma II.5.:** *There exists some constant $A$ such that:*

$$v(z) - zX'(z)^2 \geq Az \;, z \in R^+.$$

*Moreover:*

$$X_n = X_P(n) + O(1)$$
$$V_n = V_P(n) - nX_P'(n)^2 + O(1).$$

*where $X_n$ and $v_n$ are respectively the mean and the variance of the Bernoulli process.*

**Proof:** The proof of the first assertion is established in appendix. The asymptotic relationship between the Bernoulli and Poisson model is proved via the use of Rice 's integrals, see [8, 16] .

We can now claim the result:

**Theorem II.6.:** *The distribution of the size of the tries, once centered and normalized, converges to the normal distribution. Moreover, the moments of any order of the centered and normalized distribution converge to the corresponding moments of the normal distribution.*

**Proof:** $P(z, e^{\frac{it}{\sigma(z)}})e^{-itX(z)} = e^{-t^2+O(\frac{z.t^3}{\sigma^3(z)})}$. As: $|v(z)|>A. |z|$ , we have: $\frac{z}{\sigma(z)^3} < \frac{B}{\sqrt{|z|}}$. Thus: $P(z, e^{\frac{it}{\sigma(z)}})e^{-itX(z)} \longrightarrow e^{-\frac{t^2}{2}}$, uniformly in any neighbourhood of $t=0$ (as asymptotically we have $e^{it} \in V(1)$).

# III. A SEMI TAUBERIAN THEOREM DESIGNED FOR POISSON GENERATING FUNCTIONS

## III.1. INTRODUCTION.

### III.1.1. The goals.

We want to show the gaussian behaviour of the distribution of the size of tries when the conditionnal number of inserted keys indefinitely increases. The characteristic function of this distribution is $P_n(e^{-it})$ and we have the relation:

$$P(z,e^{-it}) = \sum_{n \in N} P_n(e^{-it}) \frac{z^n}{n!} e^{-z} .$$

We will essentially show the evaluation:

$$P_n(\exp(-\frac{it}{\sqrt{n}})) \sim \frac{1}{i} \int_{n-i\infty}^{n+i\infty} P(z,\exp(-\frac{it}{\sqrt{n}}))\exp(\frac{(z-n)^2}{2n}) . \frac{1}{\sqrt{2\pi n}} dz .$$

and use the asymptotics about $P(z,e^{-it})$, introduced in the first section in an adaptation of the *Central Limit* theorem. But the previous evaluation can be extracted from a more general ***semi tauberian*** theorem which binds asymptotics of complex sequences $\{a_n\}_{n \in N}$ with asymptotics of their associated Poisson generating function $a(z)$, defined by:

$$a(z) = \sum_{n \in N} a_n \frac{z^n}{n!} e^{-z} .$$

Under certain conditions of regularity of the sequence $\{a_n\}_{n \in N}$ we will show the evaluation:

$$a_n \sim \frac{1}{i} \int_{n-i\infty}^{n+i\infty} a(z)\exp\{\frac{(z-n)^2}{2n}\} . \frac{1}{\sqrt{2\pi n}} dz$$

In fact this section deals with two targets. First we develop the semi-tauberian theorem. Second we use it as a tool for the proof of the Central Limit Theorem about the size of tries when the Bernoulli model is used.

This last theorem is of course the main result of this section as far as this paper deals with asymptotics of trie parameters --and especially with the size of tries. But we must point out the importance of the first theorem which ranges over the larger field of application of the *Poisson generating* functions. For example, the section devoted to height of tries will be concluded with the use of this tauberian-like theorem. There are many other applications in numbering theory of graphs, of permutations, etc.

However we will keep the following ***plan***: the Central Limit theorem is the ***goal*** and the Semi-Tauberian theorem is the ***tool***. Now we are ready to claim

**Theorem (III-0)**: *the distribution of the random variable associated to the size of the tries, conditioned by the number n of insertions (or the number of packets in initial collision, as far as we consider the "protocol" view of the problem), converges to a normal law when n indefinitely increases.*

The analysis is based on the deduction of the properties of $P_n(e^{-it})$ from the properties of $P(z,e^{-it})$, as a complex variable function, according to the relation:

$$P(z,e^{-it}) = \sum_{n\in N} P_n(e^{-it})\frac{z^n}{n!}e^{-z} .$$

We claim the following proposition:

**Theorem (III-1):** *For t real we have the evaluation*

$$P_n(\exp(-\frac{it}{\sqrt{n}})) = \frac{1}{i}\int_{n-i\infty}^{n+i\infty} P(z,\exp(-\frac{it}{\sqrt{n}}))\exp(\frac{(z-n)^2}{2n}).\frac{1}{\sqrt{2\pi n}}dz + \varepsilon_n(t)$$

*with $\varepsilon_n(t)\to 0$ when $n\to 0$, the convergence being uniform in t on every arbitrary compact set.*

It will follow the evaluation:

**Theorem (III-0-bis):** *for t located in an arbitrarily real compact neighbourhood of 0 we have the convergence:*

$$P_n(\exp(-\frac{it}{\sqrt{n}})) = \exp\{-\frac{iX(n)t}{\sqrt{n}}-(\frac{v(n)}{n}-(X(n))^2)\frac{t^2}{2}\} + O(\frac{t}{\sqrt{n}}) .$$

This will achieve the demonstration of theorem (III-0).

**III.1.2. The tools.**

The theorem (III-1) is a corollary to a general *semi tauberian* theorem designed for *Poisson* generating functions.

Let $\{a_n\}_{n\in N}$ be a complex sequence, we call its **Poisson Generating Function** the complex function formally defined by:

$$a(z) = \sum_{n\in N} a_n \frac{z^n}{n!}e^{-z}$$

**Remark:** This definition is analogous to the definition of the **Ordinary Generating Function**, $f(z)$, defined by:

$$f(z) = \sum_{n\in N} a_n z^n(1-z) .$$

These series will be of no object here.

When $\{a_n\}_{n\in N}$ satisfies the inequation $|a_n| \le 1$ for every integer $n$, and provided certain other conditions of regularity, the semi-tauberian theorem claims the convergence:

$$\left| a_n - \frac{1}{i}\int_{n-i\infty}^{n+i\infty} a(z)\exp\{\frac{(z-n)^2}{2n}\}.\frac{1}{\sqrt{2\pi n}}dz \right| \to 0 ,$$

when $n\to 0$ (III-3), independently of the sequence $\{a_n\}_{n\in N}$. We deduce (III-1) by simple identification:

$$\{a_n\}_{n\in N} = \{P_n(\exp(-\frac{it}{\sqrt{m}}))\}_{n\in N} .$$

The asymptotical evaluation (III-0-bis) is with the results of the analysis of the Poisson process in the previous section.

**Remark:** the reason that makes us characterize this theorem as a semi-tauberian one is in the fact that we need expansions of $a(z)$ on a large part of the complex plan for deducing an evaluation of $a_n$. In a tauberian theorem, for example the theorem designed for ordinary generating functions,

$$f(1-\frac{1}{n}) \sim a_n ,$$

we need only knowledge of $f(z)$ on a real segment.

**III.1.3. The plan.**

In the sub-section II we study a new functional transform: $Z : f \to Z[f]$ with $t \in R^+$:

$$Z[f(z);t] = \frac{1}{i}\int_{t-i\infty}^{t+i\infty} f(z)\exp(\frac{(z-t)^2}{2t}).\frac{1}{\sqrt{2\pi t}}dz \qquad \text{(III-2)}$$

defined on the space of functions $f$ of exponential type.

In the sub-section III we prove the general semi tauberian theorem. We show (III-1) in the sub-section IV and achieve (III-0-bis) and (III-0) in sub-section V.

## III.2. THE TRANSFORM Z

### III.2.1. Definition

Let $f$ be a holomorph function of exponential type. *I.e.* there exist $A$ and $B$, non negative reals, such that:

$$\forall z \in C \quad |f(z)| \leq Be^{A|z|} .$$

First we define $Z[f]$ as a positive variable function:

$$t>0 : \; Z[f](t) = \frac{1}{i}\int_{t-i\infty}^{t+i\infty} f(z)\exp(\frac{(z-t)^2}{2t})\frac{1}{\sqrt{2\pi t}}dz \; . \quad (III-2)$$

Note that the integration is done along a vertical axis, and is convergent because $f$ is of exponential type.

**III.2.2. Proposition:** *Z transforms a function of exponential type into a real positive variable function also of exponential type. Furthermore this function can be analytically continued into the whole complex plane to a holomorph function of exponential type.*

**Proof:** $f$ owns a *right-side* Laplace transform $f^*$

$$f^*(\omega) = \int_0^\infty f(z)e^{-i\omega z}dz$$

which is holomorph in a neighbourhood of infinity. We know that

$$f(z) = \frac{1}{2i\pi}\oint f^*(\omega)e^{i\omega z}d\omega$$

the integration being done around a finite contour included in the neighbourhood of infinity and including all singularities of $f^*$. Thus we can write:

$$Z[f](t) = \frac{1}{i}\int_{t-i\infty}^{t+\infty}\left(\frac{1}{2i\pi}\oint f^*(\omega)e^{i\omega z}d\omega\right)\exp\left(\frac{(z-t)^2}{2t}\right)\frac{1}{\sqrt{2\pi t}}dz$$

and, permuting the integration symbols, we obtain:

$$Z[f](t) = \frac{1}{2i\pi}\oint f^*(\omega)e^{t(i\omega+\omega^2/2)}d\omega\ . \tag{III-3}$$

With this last writing it is quite obvious that $Z[f]$ is of exponential type and owns an analytic continuation with the same property by taking (III-3) with $t$ complex.

And we reach the following proposition.

**III.2.3. Proposition:** *$Z[f]$ has the right-side Laplace transform $Z^*[f]$:*

$$Z^*[f](\omega') = (f^*(\omega) - f^*(-2i-\omega))\frac{d\omega}{d\omega'}\ , \tag{III-4}$$

*with $i\omega' = i\omega+\omega^2/2$.*

**Proof:** $Z[f]$ is a holomorph function of exponential type, so $Z^*[f]$ exists and is holomorph in a neighbourhood of infinity. Thus this is sufficient to show (III-4) on an open area connected at infinity, for example for $\omega'$ such that $\mathrm{Im}(\omega')$ be less than a given constant.

At first we will point out an appropriate value for such a constant.

Let us consider (III-2). Let $K$ be the greatest value reached by $|i\omega+\omega^2/2|$, when $\omega$ describes the contour of integration. Let us take $K^*>K$. We suppose $K^*$ great enough that, when $\mathrm{Im}(\omega')<-K^*$, then the two roots $\omega_1$ et $\omega_2$ of the equation $i\omega' = ix+x^2/2$ (with unknown $x$) are different and have a norm greater than $K$ (and so they are outside the primarily contour of (III-2)).

Now we set $\mathrm{Im}(\omega') < -K^*$.

We have

$$Z[f](t) = \int_0^\infty\left(\frac{1}{2i\pi}\oint f^*(\omega)e^{t(i\omega+\omega^2/2)}d\omega\right)dt\ .$$

And permuting the symbols of integration:

$$Z[f](t) = \frac{1}{2i\pi}\oint f^*(\omega)\frac{d\omega}{i\omega'-(i\omega+\frac{\omega^2}{2})}\ .$$

The function $f^*(\omega)/(i\omega'-i\omega-\omega^2/2)$ is a holomorph function of $\omega$ with two simple poles $\omega_1$ and $\omega_2$ which are outside the contour of integration. $f^*(\omega)$, as a right-side Laplace transform, decreases at least as $\omega^{-1}$ when $|\omega|\to\infty$, so $f^*(\omega)/(i\omega'-i\omega-\omega^2/2)$ decreases faster than $\omega^{-2}$. Thus we have the equality, according to the theorem of the residues for meromorph functions:

$$Z^{*}[f](\omega') = -Residue\ in\ \omega_1 - Residue\ in\ \omega_2\,,$$

namely,

$$Z^{*}[f](\omega') = f^{*}(\omega_1)\frac{d\omega_1}{d\omega'} + f^{*}(\omega_2)\frac{d\omega_2}{d\omega'}\ .$$

With $\omega_1+\omega_2 = -2i$, the proof is achieved. ■

## III.3. A SEMI TAUBERIAN THEOREM

### III.3.1. Notation and definition

(i) Let $\varphi$ be a real $C^\infty$ function with compact support ($\varphi$ is non zero only on a bounded interval) and $t$ be a non negative real. We note

$$\varphi_t : x \to \frac{1}{\sqrt{t}}\varphi\left(\frac{x-t}{\sqrt{t}}\right);$$

$\varphi$ is considered like a *window* which slides on the real axis and which can be centered on every value $t$ with a scale dilatation $\sqrt{t}$.

(ii) We note $f_m$, $m$ integer, the complex function

$$f_m(z) = \frac{z^m}{m!}e^{-z}\ .$$

(iii) Let $\{a_n\}_{n\in N}$ be a complex sequence such that $|a_n| \le 1$ for all integer $n$. Let

$$a(z) = \sum_{n\in N} a_n \frac{z^n}{n!}e^{-z} = \sum_{n\in N} a_n f_n(z)$$

be its *Poisson function* and let us note $\Delta$ the discrete distribution

$$\Delta = \sum_{n\in N} a_n \delta_n$$

where $\delta_n$ is the *Dirac* measure at point $n$.

(iv) $\varphi$ always being a compact support function, we note

$$<\Delta|\varphi> = \sum_{n\in N} a_n\varphi(n)\ ,$$

and we extend this notation, if $f$ is a real function,

$$<f\,|\varphi> = \int_{-\infty}^{+\infty} f(x)\varphi(x)dx\ .$$

In the following we will often assume $f$ analytical and, analogously with the right-side Laplace transform, we will limit the integration:

$$<f\,|\varphi> = \int_{0}^{\infty} f(x)\varphi(x)dx\ .$$

### III.3.2. The window theorem

**Theorem (III-5):** *$\varphi$ being a given function with compact support, for every sequence $\{a_n\}_{n\in N}$, satisfying the rule (iii), we have the convergence*

$$(<Z[a]|\varphi_n> - <\Delta|\varphi_n>) \to 0$$

*when $n \to \infty$. This convergence is independent of coefficients $\{a_n\}_{n\in N}$; i.e. there exists an upper bounding sequence $\{\varepsilon_n(\varphi)\}_{n\in N}$ (only dependent on $\varphi$) such that*

$$|<Z[a]|\varphi_n> - <\Delta|\varphi_n>| < \varepsilon_n(\varphi)$$

*and $\varepsilon_n(\varphi) \to 0$ when $n \to \infty$.*

In order to show the theorem we formally separate the sum for an arbitrary $\alpha$

$$<Z[a]|\varphi_n> = <C[a]|\varphi_n> + <R[a]|\varphi_n> ,$$

$$<\Delta|\varphi_n> = <C\Delta|\varphi_n> + <R\Delta|\varphi_n> ,$$

with

$$C[a] = \sum_{|m-n|<\alpha\sqrt{n}\log n} a_m f_m ,$$

$$R[a] = \sum_{|m-n|\geq\alpha\sqrt{n}\log n} a_m f_m ,$$

and

$$C\Delta = \sum_{|m-n|<\alpha\sqrt{n}\log n} a_m \delta_m ,$$

$$R\Delta = \sum_{|m-n|\geq\alpha\sqrt{n}\log n} a_m \delta_m .$$

With the *Central Convergence Theorem* we will show

$$(<C[a]|\varphi_n> - <C\Delta|\varphi_n>) \to 0$$

when $n \to \infty$, this convergence being independent of the sequence $\{a_n\}_{n\in N}$, then with the *negligible remaining terms theorem* we will show, always independently of the sequence $\{a_n\}_{n\in N}$,

$$<R[a]|\varphi_n> \text{ et } <R\Delta|\varphi_n> \to 0 .$$

**Lemma (III-6):** *$f_m$ has the right-side Laplace transform:*

$$f_m^{\bullet}(\omega) = \frac{1}{(1+i\omega)^{m+1}} .$$

**Proof:** obvious with successive integrations by parts.

### III.3.3. central convergence theorem

**Theorem (III-7):** *Let $\alpha>0$ real and $m\in[n-\alpha\sqrt{n}\log n \,,\, n+\alpha\sqrt{n}\log n]$ integer be arbitrarily choosen, then*

$$\sqrt{n}\log n(<Z[f_m]|\varphi_n> - \varphi_n(m)) \to 0$$

*when* $n \to \infty$.

**Proof:** according to theorem (III-4) $Z[f_m]$ has the right-side Laplace transform

$$Z^*[f_m](\omega') = \left(\frac{1}{(1+i\omega)^{m+1}} - \frac{1}{(3-i\omega)^{m+1}}\right)\frac{d\omega}{d\omega'}$$

with $i\omega = 1 - \sqrt{1-2i\omega'}$.

The Fourier transform of $\varphi_n$ is:

$$\varphi_n^*(\omega) = e^{-in\omega}\varphi^*(\sqrt{n}\,\omega)$$

where $\varphi^*$ is the Fourier transform of $\varphi$.

The Parseval theorem ensures:

$$<Z[f_m]\,|\,\varphi_n> = \frac{1}{2i\pi}\int_{-\infty}^{+\infty} Z^*[f_m](-\omega')\varphi_n^*(\omega')d\omega'$$

$$= \frac{1}{2i\pi}\int_{-\infty}^{+\infty}\varphi_n^*(\omega')\left(\frac{1}{(1+(1-\sqrt{1+2i\omega'}))^{m+1}} - \frac{1}{(3-(1-\sqrt{1+2i\omega'}))^{m+1}}\right)\frac{d\omega'}{\sqrt{1+2i\omega'}}\,.$$

it is easy to show that the imaginary part $3-(1-\sqrt{1+2i\omega'})$ be less than -2 when every real value is affected to $\omega'$. Thus

$$<Z[f_m]\,|\,\varphi_n> = \frac{1}{2i\pi}\int_{-\infty}^{+\infty}\varphi_n^*(\omega')(1+(1-\sqrt{1+2i\omega'}))^{-(m+1)}\frac{d\omega'}{\sqrt{1+2i\omega'}} + O(2^{-n})\,.$$

It is also easy to show that $2-\sqrt{1+2i\omega'}$ be ouside the unit disk. Thus we can divide the integration and write for arbitrary $B>0$, with an obvious change of variable:

$$<Z[f_m]\,|\,\varphi_n> = \frac{1}{2i\pi}\frac{1}{\sqrt{n}}\int_{-Bn^{1/6}}^{+Bn^{1/6}}(1+(1-(1-\frac{2i\omega'}{\sqrt{n}})^{1/2}))^{-(m+1)}\cdot\frac{e^{-i\omega'\sqrt{n}}\varphi^*(\omega')}{(1+\frac{2i\omega'}{\sqrt{n}})^{1/2}}d\omega'$$

$$+\ O\left[\frac{1}{\sqrt{n}}\int_{R-[-Bn^{1/6},Bn^{1/6}]}|\varphi(\omega')|\,d\omega'\right]. \qquad \text{(III-8)}$$

We expand

$$(1+(1-(1+\frac{2i\omega'}{\sqrt{n}})^{1/2}))^{-(m+1)} = \exp\{-(m+1)(-\frac{i\omega'}{\sqrt{n}} + O(\frac{\omega'^3}{n^{3/2}}))\}$$

$$= \exp\{m\frac{i\omega'}{\sqrt{n}} + O(\frac{\omega'^3}{\sqrt{n}})\}\,.$$

This expansion is valid because $\omega'/\sqrt{n}$ is confined in a compact set. Indeed $\omega'^3/\sqrt{n}$ is upper bounded by $B$; precisely for this reason we can extract it from the exponential and then establish the general evaluation:

$$(1+\frac{2i\omega'}{\sqrt{n}})^{-1/2}\,(1+(1-(1+\frac{2i\omega'}{\sqrt{n}})^{1/2})^{-(m+1)} = \exp\{m\frac{i\omega'}{\sqrt{n}}\}(1 + O(\frac{q_3(|\omega'|)}{\sqrt{n}}))$$

where $q_3$ is a positive coefficient polynomial of degree 3. We can insert these evaluations in (III-8)

$$<Z[f_m]|\varphi_n> = \frac{1}{2i\pi}\frac{1}{\sqrt{n}}\int_{-Bn^{1/6}}^{+Bn^{1/6}} \exp(m\frac{i\omega'}{\sqrt{n}})e^{-i\omega'\sqrt{n}}\varphi^*(\omega')d\omega'$$

$$+ O\left(\int_{R-[-Bn^{\frac{1}{6}},Bn^{\frac{1}{6}}]} |\varphi(\omega)|d\omega\right) + \frac{1}{n}O\left(\int_{-\infty}^{+\infty} q_3(|\omega|)|\varphi^*(\omega)|d\omega\right).$$

The last term exists because $\varphi^*$ is $C^\infty$ fast decreasing as the Fourier transform of a $C^\infty$ fast decreasing function (a $C^\infty$ fast decreasing function g is such that $\forall\, k,l \in N$: $x^k g^{(l)}(x) \to 0$ when $|x| \to \infty$). According to this property we also have, for every integer $k$:

$$\int_{R-[-Bn^{\frac{1}{6}},Bn^{\frac{1}{6}}]} |\varphi(\omega)|d\omega = O(n^{-k})$$

and

$$\frac{1}{2i\pi}\frac{1}{\sqrt{n}}\int_{-B\sqrt{n}}^{+B\sqrt{n}} \exp(m\frac{i\omega'}{\sqrt{n}})e^{-i\omega'\sqrt{n}}\varphi^*(\omega')d\omega' = \frac{1}{2i\pi}\frac{1}{\sqrt{n}}\int_{-\infty}^{+\infty} \exp(\frac{m-n}{\sqrt{n}}i\omega)\varphi^*(\omega)d\omega$$

$$+ O(n^{-k})$$

$$= \varphi_n(m) + O(n^{-k}) .$$

The last expression is deduced from the classical Fourier inversion. We can write:

$$\sqrt{n}\,\log n\,(<Z[f_m]|\varphi_n> - \varphi_n(m)) = \frac{\log n}{\sqrt{n}}O(1) .$$

This achieves the proof of the theorem. ■

### III.3.4. The Negligible Remaining Terms Theorem

**Theorem (III-9):** *Let $t$ belong to the support of $\varphi_n$ (of the form $[n-\beta\sqrt{n}, n+\beta\sqrt{n}]$, $\beta$ being a positive real), let $m$ be such that $|m-n|>\alpha\sqrt{n}\log n$, then for every integer $k$, we have:*

$$Z[f_m](t) = O((m-t)^{-k})$$

*when $n \to \infty$.*

**Proof:** the proof is enclosed in appendix. We can draw a scheme.

- We come back to the expression (III-3)

$$Z[f_m](t) = \frac{1}{2i\pi}\oint\frac{1}{(1+i\omega)^{m+1}}e^{t(i\omega+\omega^2/2)}d\omega ,$$

  integrated along a contour around $i$. We choose the contour in such a way that it includes 0 and that $|1+i\omega|^{-1}$ and $|exp\{i\omega+\omega^2/2\}|$ be less than 1.

- We expand as in theorem (III-7):

$$(1+i\omega)^{-m-1} = \exp\{-(m+1)(i\omega+\frac{\omega^2}{2}-\frac{i\omega^3}{3}-\frac{\omega^4}{4}+O(\omega^5))\} ,$$

$$(1+i\omega)^{-m} e^{t(i\omega+\omega^2/2)} = \exp\{(m-t)(i\omega+\frac{\omega^2}{2})-m(\frac{i\omega^3}{3}+\frac{\omega^4}{4}+O(\omega^5))\} .$$

- Through a suitable parametrisation we reach the evaluation:

$$Z[f_{m-1}](t) = \frac{v}{2i\pi}\left[\frac{|t-m|}{m+t}\right]^{1/2}\int_{-\infty}^{+\infty}\exp\{i\xi(\tau'' + v\frac{\tau''^3}{6})\}g_{m,t}(\eta\tau'')d\tau''$$

$$+ O(D^{-m-t})$$

with

$$\eta = \left[\frac{|t-m|}{(m+t)^{1/2}}\right]^{1/2} ,$$

$$v = \operatorname{sign}(t-m) ,$$

$$\xi = v\left[\frac{|t-m|^3}{m+t}\right]^{1/2} .$$

$D > 1$

$$g_{m,t}(\tau) = e^{-\frac{17}{48}\tau^4}(1+O(\frac{\tau^5}{(m+t)^{1/4}})) .$$

- We achieve the proof, with $\xi$ et $\eta \to\infty$ when $n\to\infty$ and an application of the asymptotical expansions provided by the ***stationary phase theorem.***

The theorems (III-7) and (III-9) prove the window theorem. ■

## III.4. CONDITIONS OF REGULARITY

The goal of this sub section is to establish some sufficient conditions for the convergence

$$(a_n - Z[a](n)) \to 0$$

when $n\to\infty$ (III-10), as issue from the window theorem.

### III.4.1. Definition

We will say that a double sequence $\{a_n^h\}_{n,h\in N}$ is *locally normally varying* when for every positive real $\beta$ there exists $\gamma_d$ positive real such that for every integer $n$ and $m\in[n-\beta\sqrt{n}\,,\,n+\beta\sqrt{n}\,]$ we can upper-bound

$$|a_m^n - a_n^n| \le \gamma_d\,\frac{|m-n|}{\sqrt{n}} .$$

According to the same idea, we will say that a sequence of functions with real variable $x$ $\{f^h(x)\}_{h\in N}$ is also locally normally varying when for every positive real $\beta$ there exists $\gamma_c$ positive real such that for every integer $n$ and real $x \in [n-\beta\sqrt{n}\,,\, n+\beta\sqrt{n}\,]$ we can upper-bound

$$|f^n(x) - f^n(n)| \le \gamma_c \frac{|x-n|}{\sqrt{n}} .$$

In the following we set the relation, with $z$ complex and $h$ integer,

$$a^h(z) = \sum_{n\in N} a_n^h \frac{z^n}{n!} e^{-z} .$$

### III.4.2. Normal variation and monotonous variation theorems

**Theorem (III-11):** *Let $\{a_n^h\}_{n,h\in N}$ be a double sequence satisfying the condition $|a_n^h| < 1$ for every integer $n$ and $h$. When $\{a_n^h\}_{n,h\in N}$ and $\{Z[a^h](x)\}_{h\in N}$ are locally normally varying then we have the convergence*

$$(a_n^n - Z[a^n](n)) \to 0 \qquad \textit{(III-10)}$$

*when $n\to\infty$. Moreover this convergence is uniform on every class of double sequence $\{a_n^h\}_{n,h\in N}$ for which $\gamma_d$ and $\gamma_c$ are bounded.*

**Proof:** let us take a *test* function $\varphi$ with compact support:

$$<\Delta^n | \varphi_n> = \sum_{m\in N} a_m^n \varphi_n(m) = a_n^n\Big(\sum_{m\in N} \varphi_n(m)\Big) + O\left(\frac{\sum_{m\in N} |(m-n)\varphi_n(m)|}{\sqrt{n}}\right).$$

According to Riemann we have

$$\sum_{m\in N} \varphi_n(m) = \int\varphi(x)dx + O\Big(\frac{1}{\sqrt{n}}\Big) .$$

Thus

$$<\Delta^n | \varphi_n> = a_n^n \int\varphi(x)dx + O\left(\int |x\,\varphi(x)|\,dx + \frac{1}{\sqrt{n}}\right).$$

In the same way

$$\begin{aligned} <Z[a^n] | \varphi_n> &= \int Z[a^n](x)\varphi_n(x)dx \\ &= Z[a^n](n)\int\varphi_n(x)dx + \int (Z[a^n](x) - Z[a^n](n))\varphi_n(x)dx \;, \\ &= Z[a^n](n)\int\varphi(x)dx + O\Big(\int |x\varphi(x)|\,dx\Big) . \end{aligned}$$

Then if we choose $\varphi$ such that $\int\varphi(x)dx = 1$ and $\int |x\varphi(x)|\,dx$ arbitrarily little we prove the theorem. ■

**Theorem (III-12):** *in order to maintain the convergence (III-10) it is sufficient that $\{a_n^h\}_{n,h\in N}$ be a real positive monotonous sequence with respect to index $n$ and $\{Z[a^h](x)\}_{h\in N}$ locally normally varying. This convergence will be independent of coefficients $\{a_n^h\}_{n,h\in N}$ assuming a bounded $\gamma_c$.*

**Proof:** we use the same proof as before with the slight difference that for evaluating $a_n^n$ we take an under- and upper-bounding with two test positive functions $\varphi^+$ and $\varphi^-$ of support respectively included in $R^+$ and $R^-$. We presume that $\{a_n^h\}_{n,h\in N}$ be increasing with respect to $n$. Thus we range:

$$<\Delta^n \mid \varphi_n^-> \leq a_n^n \sum_{m\in N} \varphi_n^-(m) = a_n^n(\int \varphi^-(x)dx + O(\frac{1}{\sqrt{n}}))$$

and

$$<\Delta^n \mid \varphi_n^+> \geq a_n^n \sum_{m\in N} \varphi_n^+(m) = a_n^n(\int \varphi^+(x)dx + O(\frac{1}{\sqrt{n}}))$$

Thus if we take $\int \varphi^-(x)dx = \int \varphi^+(x)dx = 1$ and $\int |x\varphi^\pm(x)|dx$ arbitrarily little we prove this new theorem. ■

## III.5. EVALUATION OF $P_n(e^{-i\frac{t}{\sqrt{n}}})$, $t\in R$

### III.5.1. Conditions of regularity

We show in appendix that the characteristic function of the distribution of size of trie, under the Bernoulli model, has the uniform evaluation:

$$P_n(e^{-it}) = O(t)$$

uniformly with respect to $n$ and $t$ confined in a real compact neighbourhood of 0. Thus the sequence $\{P_n(e^{-i\frac{t}{\sqrt{h}}})\}_{n,h\in N}$ is locally normally varying. This property is to be shown within the sequence of functions $\{Z[P(z,e^{-i\frac{t}{\sqrt{h}}});x]\}_{h\in N}$:

**Theorem (III-13):** *for $x$ in an interval like $[n-\beta\sqrt{n}, n+\beta\sqrt{n}]$, $\beta$ real positive, we have the uniform convergence with respect to $t$ in an arbitrarily compact neighbourhood of 0:*

$$Z[P(z,e^{-i\frac{t}{\sqrt{n}}});x] - \exp\{-X(n)\frac{it}{\sqrt{n}} - \frac{(v(n)-nX'^2(n))}{2n}t^2\} \to 0$$

*when $n\to\infty$.*

We need the following lemma:

**Lemma (III-14):** *we can truncate the integration, for an arbitrary $\alpha>0$ and $x>0$,*

$$Z[a](x) = T[a](x) + \varepsilon(x)$$

*with*

$$T[a](x) = \frac{1}{i}\int_{x-\alpha ix^{3/4}\log x}^{x+\alpha ix^{3/4}\log x} a(z)\exp\{\frac{(z-x)^2}{2x}\}\frac{dz}{\sqrt{2\pi x}},$$

*and $\varepsilon(x)\to 0$ when $x\to\infty$, uniformly for all sequences $\{a_n\}_{n\in N}$ satisfying the condition III 1 (iii).*

**Proof:** we use the general inequality $|a(z)| \le e^{|z|-\mathrm{Re}(z)}$, Re$(z)$ describing the real part of the complex number $z$. Then

$$Z[a](x) - T[a](x) = O\left[\int_{\alpha x^{1/4}\log x}^{\infty} \exp\{\sqrt{x^2+\vartheta^2 x} - \frac{\vartheta^2}{2} - x\}d\vartheta\right].$$

Now using either the expansion of $\sqrt{1+\frac{\vartheta^2}{x}}$ near 0 when, for example, $\vartheta \le 3\sqrt{x}$:

$$\sqrt{1+\frac{\vartheta^2}{x}} - 1 + \frac{\vartheta^2}{2x} < -A\frac{\vartheta^4}{x^2}$$

($A$ being a nonnegative real), or the trivial inequality:

$$\sqrt{x^2+\vartheta^2 x} < x + \vartheta\sqrt{x}\ ,$$

when $\vartheta > 3\sqrt{x}$, one gets the upper bound of the integrand:

$$\begin{cases} \sqrt{x^2+x\vartheta^2} - \dfrac{\vartheta^2}{2} - x \le -A\dfrac{\vartheta^4}{x} \le -A(\alpha\log x)^4 \\ \sqrt{x^2+x\vartheta^2} - \dfrac{\vartheta^2}{2} - x < -\vartheta(\dfrac{\vartheta}{2} - \sqrt{x}) < -\dfrac{\sqrt{x}\,\vartheta}{2} \end{cases}$$

Integrating steadily gives:

$$Z[a](x) = T[a](x) + O(\sqrt{x}e^{-A(\alpha\log x)^4} + \frac{e^{-\frac{3}{2}x}}{\sqrt{x}})\ .$$

■

**Proof of theorem (III-13):** we recall the following asymptotical expressions, when $n\to\infty$ (then $x\to\infty$). These expressions are established and expanded in the section II, so we just mention their conditions of existence.. $\vartheta$ is located in an interval such as $[-\gamma\sqrt{x}\ ,\gamma\sqrt{x}\ ]$, with $\gamma$ real positive, in order that $x+i\vartheta\sqrt{x}$ be maintained in a cone. Moreover we assume that $\exp\{it/\sqrt{n}\}$ in the compact neighbourhood of 1 mentioned in the first section (for $t$ in an arbitrary compact set this condition is of course satisfied as $n\to\infty$).

$$P(x+i\vartheta\sqrt{x}\ , e^{-i\frac{t}{\sqrt{n}}}) =$$
$$\exp\{-X(x+i\vartheta\sqrt{x}\ )\frac{it}{\sqrt{n}} - v(x+i\vartheta\sqrt{x}\ )\frac{t^2}{2n} + O(t^3\frac{(x+i\vartheta\sqrt{x}\ )}{n^{3/2}})\}$$

Always according to the asymptotical analysis we get the evaluations

$$X(x+i\vartheta\sqrt{x}\ ) = X(x) + X'(x)i\sqrt{x}\,\vartheta + O(\frac{1}{x}(\sqrt{x}\,\vartheta)^2)\ ,$$

since $X(z) = O(z)$, $X'(z) = O(1)$, $X''(z) = O(1/z)$, for $z\to\infty$ in a complex cone. In the same way

$$v(x+i\vartheta\sqrt{x}) = v(x) + O(\sqrt{x}\,\vartheta)\,,$$

because $v(z) = O(z)$, $v'(z) = O(1)$. Thus we have the general evaluation:

$$P(x+i\vartheta\sqrt{x}\,,\, e^{-i\frac{t}{\sqrt{n}}}) = \exp\{-X(x)\frac{it}{\sqrt{n}} + X'(x)\frac{\sqrt{x}}{\sqrt{n}}t\vartheta - v(x)\frac{t^2}{2n}$$

$$+ O(\frac{\vartheta^2 t}{\sqrt{n}}) + O(\frac{\vartheta t^2}{\sqrt{n}}) + O(\frac{t^3(1+\vartheta/\sqrt{n})}{\sqrt{n}})\}\,.$$

Thus when $n \to \infty$, with $\gamma = x^{\frac{1}{4}}\log x$ the expression,

$$T[P(z,e^{\frac{-it}{\sqrt{n}}});x] = \frac{1}{\sqrt{2\pi}}\int_{-\gamma\sqrt{x}}^{\gamma\sqrt{x}} P(x+i\vartheta\sqrt{x}\,,\, e^{-i\frac{t}{\sqrt{n}}})e^{-\frac{\vartheta^2}{2}}d\vartheta\,,$$

tends to, as $\gamma\sqrt{x} \to \infty$:

$$\frac{1}{\sqrt{2\pi}}\exp\{-X(x)\frac{it}{\sqrt{n}} - v(x)\frac{t^2}{2n}\}\int_{-\infty}^{+\infty} e^{X'(x)t\frac{\sqrt{x}}{\sqrt{n}}\vartheta - \frac{\vartheta^2}{2}}d\vartheta\,;$$

namely:

$$\exp\{X(x)\frac{it}{\sqrt{n}} - (v(x)-xX'^2(x))\frac{t^2}{2n}\}\,.$$

Thus with the lemma (III-13) the theorem is shown. ■

### III.5.2. Normal convergence of the distribution of the size of the tries

We achieve here the proof of the theorem (III-0).

According to the previous analysis and the asymptotical results within $X(z)$ and $v(z)$ it is obvious that $\{Z[P(z,\exp\{-it/\sqrt{n}\});x]\}_{n\in N}$ be a sequence of functions,of variable $x$, locally normally varying. Consequently the hypotheses of the semi tauberian theorem are satisfied and we can claim:

$$P_n(e^{-\frac{it}{\sqrt{n}}}) - \exp\{X(x)\frac{it}{\sqrt{n}} - (v(x)-xX'^2(x))\frac{t^2}{2n}\} \to 0$$

when $n \to \infty$, and $t$ remain in an arbitrary real compact neighbourhood of 0.

According to the first section we know that

$$X_n = X(n) + O(1)$$

and

$$v_n = v(n) - nX'^2(n) + O(1) = \omega(n)$$

($v_n$ increases at least as a linear function of $n$). Thus the normalized and centered law of the size of trie conditioned by the number $n$ of keys gets the generating function

$$e^{iX_n\frac{t}{\sqrt{v_n}}}P_n\left(e^{-\frac{it}{\sqrt{v_n}}}\right)$$

which is evaluated, for $t$ in an arbitrary compact real neighbourhood of 0, by

$$\exp\{O(\frac{it}{\sqrt{v_n}}) - \frac{v(n) - nX^2(n)}{v_n}\frac{t^2}{2}\},$$

which converges to

$$e^{-\frac{t^2}{2}}.$$

■

## IV. DEPTH OF INSERTION:

We show here how to generalize our results of previous sections. We keep on using generating functions, first studying Poisson distributions, and then translating the results to Bernoulli distributions. We are dealing here with a parameter ***depth of insertion***, defined below, which is related to the external path length. This case is of interest as it shows a new way to prove the convergence of Poisson distributions, when $P(z,u)$ satisfies some additive -or quasi-linear- equation. In the reasoning we shall emphasize the differences and analogies with Section II, reporting the details of the modified demonstrations to the Annexes or to references.

### IV.1. RECURSION AND FUNCTIONAL EQUATION

**Definition**: The **depth of a leaf** is the number of internal nodes on the path from the root to this leaf.

We consider here the distribution of this parameter. We note:

$P_n^k$ := *Prob(a record is inserted at depth k in a trie with n+1 records).*

**Remark 1**: Such a probability is defined when the trie is not empty. Thus, this shift makes our notation consistent with the previous ones, where we had sums from $n=0$.

**Remark 2**: We find that : $(n+1)P_n'(1)$ is the average path length, when we count the number of *internal* nodes from the root to every leaf in a trie with $n+1$ records.

We can state a proposition analogous to Proposition II.1 :

**Proposition IV.1**: *Let $P(z,u)$ be the generating function for the depth of the leaves in random tries. $P(z,u)$ is defined and analytical with respect to z and u in the domain $C\times B(0,\frac{1}{p^{b+1}+q^{b+1}})$. It satisfies there the functional equation:*

$$P(z,u) = u\,[pP(pz,u)+qP(qz,u)] + (1-u)e_{b-1}(z)e^{-z}, \qquad (IV.1)$$

*and the growth property* $(P1)$.

**Proof**: Defining formally $P_{-1}^k \equiv 0$, $P_{-1}(u) \equiv 0$, we have, for $n+1>b, k\geq 1$:

$$P_n^k = \sum_{\substack{n_1+n_2=n-1\\ n_1,n_2\geq -1}} \binom{n+1}{n_1+1,n_2+1} p^{n_1+1}q^{n_2+1}\left(\frac{n_1+1}{n+1}P_{n_1}^{k-1}+\frac{n_2+1}{n+1}P_{n_2}^{k-1}\right).$$

Writing $K_{n+1}(u) = (n+1)P_n(u)$, we get:

$$\begin{cases} K_{n+1}(u) = u \displaystyle\sum_{\substack{m_1+m_2=n+1\\ m_1,m_2\geq 0}} \binom{m+1}{m_1+1,m_2+1} p^{m_1}q^{m_2}(K_{m_1}(u)+K_{m_2}(u)),\ n+1>b, \\ \qquad K_0(u)=0,\ K_1(u)=\ldots=K_b(u)=1. \end{cases}$$

Thus:

$$K(z,u) = u\ [pK(pz,u)+qK(qz,u)] + (1-u)ze_{b-1}(z)e^{-z},$$

and equation $(IV.1)$ follows as $zP(z,u)=K(z,u)$. We prove by induction: $|K_n(u)| < \alpha^{n-1}$ , where $\alpha>1$ satisfies: $(p+\frac{q}{\alpha})^b+(q+\frac{p}{\alpha})^b < \frac{1}{m}$ with $m=\sup_D \frac{|u|}{|1-u(p^{b+1}+q^{b+1})|}$, and $(P1)$ follows.

We notice that $(IV.1)$ is an *additive* equation, what $(E1)$ was not. This will make possible to derive directly an analytic development of $P(z,e^{it})$, without considering $\log P(z,e^{it})$. This skips part of the previous studies and, in particular, $(P2)$ will be of no use here.

## IV.2. ASYMPTOTIC ANALYSIS, USE OF THE MELLIN TRANSFORM

To develop $P(z,u)$, we use Mellin transform techniques [8, 17] . We note $P_u^*(s)$ the Mellin transform of $P(z,u)-1$ with respect to the variable $z$, i.e.:

$$\int_0^\infty (P(z,u)-1)z^{s-1}dz = P_u^*(s)\ .$$

Transforming $(IV.1)$ as in [8] , we get:

$$P_u^*(s) = -\frac{(1-u)\Gamma(s+b)}{s(b-1)!(1-u(p^{1-s}+q^{1-s}))}\ ,$$

in the domain where the right function is analytical, and the integral above is absolutely convergent. This leads us to the study of the roots of $(IV.2)$:

$$1-u(p^{1-s}+q^{1-s}) = 0 \qquad \text{(IV.2)}$$

or, if $u \neq 0$, we can note $u = e^{it}$ and study the roots of

$$e^{-it} = p^{1-s}+q^{1-s}\ .$$

We claim:

**Theorem IV.2.:**

(1) *The set of the roots of $(IV.2)$ is a countable set $\{s_k(t);k \in Z\}$. Moreover, for every compact set for $u$ and for every strip for $s$ : $\{s: a<\mathrm{Re}(s)<b\}$, there exists a constant $\delta(p,q)$ such that:*

$$\forall (k,k'): |s_k(t)-s_{k'}(t)| \geq \delta(p,q)\ .$$

(2) *There exist a neighbourhood $V(0)$ and an analytical function: $t \longrightarrow s_0(t)$ such that $s_0(t)$ be a root of $(IV.2)$ and $s_0(t) \underset{t\to 0}{\longrightarrow} 0$ . Moreover:*

$$s_0(t) = -\frac{it}{H} + \frac{t^2}{H^3}(H_2-H^2) + O(t^3)$$

*with:*

$$H = -(p\log p+q\log q)\ ,\ H_2 = p\log^2 p+q\log^2 q\ .$$

(3) *When $p=q=\frac{1}{2}$, we have : $\{s_k(t);k \in Z\}=\{s_0(t)+\frac{2il\pi}{\ln 2};k \in Z\}$. When $p \neq q$, we have, for any $t$ and $k$ : $Re(s_k(t)) \geq s_0(i\mathrm{Im}(t))$.*

(4) *As* $|u|<\alpha=\frac{1}{p^{b+1}+q^{b+1}}$, *all roots are contained in a half plan* $\{s:\sigma_1(\alpha)<\mathrm{Re}(s)\}$. *If* $\beta<|u|<\alpha$, *there exists a strip* $\{s:\sigma_1(\alpha)<\mathrm{Re}(s)<\sigma_2(\beta)\}$ *containing for every u all the poles of (IV.2).*

**Proof**: (1) is proved in Annex. Now, (2) is a consequence of Implicit Functions Theorem, and, by taking the successive derivatives in (*IV*.2), we get the development. When $p=q=\frac{1}{2}$, the second order coefficient disappears and the roots of $2^s=e^{it}$ simply are $s_k(t)=-\frac{it}{\ln 2}+\frac{2ik\pi}{\ln 2}$. When $p\neq q$, we notice:

$p^{1-Re(s)}+q^{1-Re(s)}\geq|p^{1-s}+q^{1-s}|=|e^{-it}|=p^{1-s_0(i\mathrm{Im}(t))}+q^{1-s_0(i\mathrm{Im}(t))}$,

which establishes (3), as $s_0(i\mathrm{Im}(t))$ is real. Finally, considering the variations of $|p^{1-s}+q^{1-s}|\to 0$ when $s\to-\infty$, we get (4).

We prove now a counterpart to Theorem II.3.

**Theorem IV.3:** *There exist a compact neighbourhood* $V(0)\subset C$ *and a cone* $C_{0,\vartheta}$ *such that, for t in* $V(0)$ *and z in* $C_{0,\vartheta}$:

$$P(z,e^{it})=e^{-\frac{it}{H}\log z-\frac{t^2}{2}\frac{(H_2-H^2)}{H^3}\log z+O(t^3.\log z)}\times(1+O(t.|z|^{-At^2}).$$

**Proof**: Considering $P_u^*(s)$, one may steadily prove (see [8]):

$$P(z,e^{it})=[z^{-s_0(t)}r_0(t)+\sum_{k\neq 0}z^{-s_k(t)}r_k(t)+O(\frac{1}{z})],$$

where $r_k(t)$ is the residue in the simple pole $s_k(t)$ of $P_u^*(s)$, i.e.: $-\frac{(1-e^{it})}{s_k(t)}\frac{\Gamma(s_k(t)+b)}{(b-1)!}.\frac{1}{H}$. Using Theorem IV.2., we may upper bound the sum $\sum_{k\neq 0}$. We write:

$$\sum_{k\neq 0}|r_k(t)z^{s_0(t)-s_k(t)}|=e^{-\mathrm{Im}(s_0(t))Arg(z)}\sum_{k\neq 0}|r_k(t)e^{\mathrm{Im}(s_k(t))Arg(z)}|.||z|^{Re(s_0(t)-s_k(t))}.$$

Any finite surface $S$ of $I$ contains at most $\frac{S}{4\pi\delta^2}$ poles. Moreover, $\Gamma$ vanishes at infinity and, as $u$ is bounded, $|\Gamma(u+iv)|\leq Ce^{-\pi|v|}$. Thus, there exist some $V(0)$ and some constant $B$ such that:

$$e^{\vartheta|\mathrm{Im}(s_0(t))|}\sum_{k\neq 0}|r_k(t)e^{\vartheta|\mathrm{Im}(s_k(t))|}|<B.|1-e^{it}|.$$

From: $Re(s_k(t))\geq s_0(i\mathrm{Im}(t))$ and the development of $s_0(t)$ in 0, we get:

$$Re(s_0(t))-Re(s_k(t))\leq Re(s_0(t))-s_0(i\mathrm{Im}(t))=O(t^2)\leq -A.t^2.$$

Thus: $\sum_{k\neq 0}|r_k(t)z^{s_0(t)-s_k(t)}|<B.|z|^{-A.t^2}$. Now, the development of $s_0(t)$ in (2) and of $r_0(t)$ ensure:

$$\begin{cases} r_0(t)=1+O(t) \\ z^{-s_0(t)}=e^{-\frac{it}{H}\log z-\frac{t^2}{2}\frac{(H_2-H^2)}{H^3}\log z+O(t^3.\log z)} \end{cases}$$

We need not here a counterpart to Lemma II.4. Nevertheless, by the cumulants formula, we can derive directly from the previous result:

**Theorem IV.4.**: *The mean and the variance of the depth of insertion satisfy asymptotically, for $z \in R^+$ :*

$$X(z) \sim \frac{1}{H}\log z + O(1),$$

$$v(z) \sim \frac{H_2 - H^2}{H^3}\log z + O(1)\,,\, p \neq q\ ,$$

$$v(z) = O(1)\,,\, p = q = \frac{1}{2}\,..$$

We assume now that $p \neq q$ and prove:

**Theorem IV.5.: Convergence to the Normal Distribution when $p \neq q$**
*The Poisson distribution of the depth of insertion, once centralized and normalized, converges to the normal distribution, when $p \neq q$.*

The proof is the same as in Theorem II.4, with $\sigma(z) \sim \sqrt{\log z}$ ,$z \in R_+$, when $p \neq q$.

We are dealing now with the case $p = q = \frac{1}{2}$. We state:

**Theorem IV.6.: Periodic Behaviour of the Distribution when $p = q = \frac{1}{2}$**
*When $z \to \infty$, the generating function satisfies asymptotically:*

$$P(z, e^{it}) = G_z(t) + O(\frac{1}{z})$$

*where:*

$$G_z(t) = \frac{z^{\frac{-it}{\ln 2}}}{\ln 2}\left(\frac{\Gamma(b - \frac{it}{\ln 2})}{(b-1)!} + \sum_{k \neq 0} \frac{\Gamma(b - \frac{it - 2ik\pi}{\ln 2})}{\Gamma(b)} z^{\frac{2ik\pi}{\ln 2}}\right)$$

*and $O(.)$ is uniform w.r.t. $u$ in any closed set included in $B(0, \frac{1}{p^{b+1} + q^{b+1}})$.*
*The moments of any order are asymptotically equivalent to the moments of $G_z(t)$.*
**Proof**: As above, we use Mellin Transform as in [8] and Theorem IV.2., (4).

**Remark**: When $z \in R^+$ the generating function is the generating function of the process under Poisson hypothesis, the distribution is asymptotically periodic in $\ln z$ translated of the mean $-\ln z / \ln 2$.

## IV.3. EQUIVALENCE POISSON-BERNOULLI:

We proceed now with the Bernoulli case. To translate the Poisson results, we use here Rice 's integrals and Newton series. [8, 16 ch. 8 ].

**Lemma IV.7.**: *The Poisson and Bernoulli distribution converge. More precisely, for $u$ in a compact neighbourhood of 0 and 1 strictly included in $B(0, \frac{1}{p^{b+1} + q^{b+1}})$, we have the uniform convergence, when $n \to \infty$:*

$$P_n(u) = P(n,u) + O(n^{-s_0(i\,\mathrm{Im}(t))-1})$$

**Proof**: For any fixed $u=e^{it}$, the following formula [8, 16 ch. 8 ] holds for $\mathrm{Im}(t)>0$:

$$a_n = P_n(u) = \frac{n!}{2i\pi}\int_{A_1}\frac{P_u^*(s)}{\Gamma(s+n+1)}ds$$

where $A_1$ is some contour around $0,-1,\dots,-n,\dots$, which does not contain any pole of $P_u^*$. Such a contour can be found, as $Re(s_k(t))\geq s_0(i\,\mathrm{Im}(t))$. Note that for $\mathrm{Im}(t)\leq 0$, we have the analogue:

$$a_n = P_n(u) = \frac{n!}{2i\pi}\int_{A_2}\frac{P_u^*(s)}{\Gamma(s+n)}ds$$

where $A_2$ is around $-1,\dots,-n$.

Deforming this contour $A_1$ in a new rectangular one, with borders going to $\infty$, we get:

$$a_n(u) = \sum_k \varphi_k(t)+O(n^{-m}).$$

with $\varphi_k(t) = n!Res(\frac{P_u^*(s)}{\Gamma(s+n)},s_k(t))$. We notice that:

$$\varphi_k(t) = \frac{1-e^{it}}{H.s_k(t)}\prod_{j=b}^{n}(1+s_k\frac{(t)}{j})^{-1}.$$

Moreover:

$$\varphi_0(t) = \frac{1-e^{it}}{s_0(t)}.\frac{1}{H}exp(1+s_0(t)(H_n-H_{b-1})+O(|s_0(t)|^2)).$$

From:

$$\begin{cases} s_0(t) = -\frac{it}{H}+O(t^2) \\ X(n) = \frac{\log n}{H} + O(1) = \frac{H_n}{H} + O(1) \\ V(n) = A\log n + O(1) \end{cases}$$

we get the result:

$$P_n(e^{\frac{it}{\sigma(n)}})\, e^{-\frac{itX(n)}{\sigma(n)}} = (1+O(\frac{t}{\sqrt{\log n}}))e^{-\frac{t^2}{2}}.$$

The foregoing proves:

**Theorem IV.8:** *The convergence results of Theorems IV.5. and IV.6. still hold under Bernoulli hypotheses.*

## V. HEIGHT OF TRIES AND NUMBER OF COLLISIONS:

An other example is given by the *height of the tries* with $n$ keys. This is the largest depth of insertion. In the case of protocols modellised by tries, this is the maximal number of collisions on one packet during a session with $n$ messages. This provides a second application of the tauberian Theorem.

The originality of this section is that the derivation is based on an approximation of the Poisson generating function $H(z,u)$ by an other -simpler- generating function $F(z,u)=\sum_k F^k(z)u^k$. First, we state general properties of $H$ and $F$. Then, we approximate their difference. Finally, we establish asymptotic results on $F$ and further on $H$. And using the semi-tauberian Theorem of Section III, we can treat the Bernoulli case.

## V.1. PROPERTIES OF GENERATING FUNCTIONS:

We note here: $g_b(z) = -\log e_b(z)e^{-z} = -\log e_b(z)e^{-z}$.
and we note $H_n^k$ the probability that a trie with $n$ keys has a height less than or equal to $k$.

**Proposition V.1**: *The probability generating function* $H(z,u)$ *satisfies*:

$$H(z,u) = (1-u)\sum_{k\geq 0} H^k(z)u^k \tag{V.1}$$

$$= (u-1)\sum_{k\geq 0}(1-H^k(z))u^k + 1$$

*where* $H^k(z)$ *is the generating function for the "cumulate probabilities"*:

$$H^k(z) = \sum_{n\in N} H_n^k \frac{z^n}{n!}e^{-z}$$

$$= \prod_{k1+k2=k} f_b(zp^{k_1}q^{k_2})^{\binom{k}{k_1,k_2}}$$

$$H^k(z) = \exp\left(-\sum_{k1+k2=k}\binom{k}{k_{1},k_2} g_b(zp^{k_1}q^{k_2})\right). \tag{V.2}$$

*It is defined and analytical for* $(z,u)$ *in* $C\times B(0,\frac{1}{p^{b+1}+q^{b+1}})$.

**Proof**: Assume a trie with $n$ keys has a height less than or equal to $k$. As both subtrees, if they exist, have then a height less than or equal to $k-1$, we get, considering the bit distribution:

$$\begin{cases} H_n^k = \sum_{n_1+n_2=n}\binom{n}{n_1,n_2}p^{n_1}q^{n_2}H_{n_1}^{k-1}H_{n_2}^{k-1}, & k\geq 1 \\ H_n^0 = \chi_{n\leq b} \end{cases}$$

Thus, $H^k(z)=\sum_n H_n^k\frac{z^n}{n!}e^{-z}$ satisfies the functional equations:

$$\begin{cases} H^k(z) = H^{k-1}(pz).H^{k-1}(qz) \\ H^0(z) = f_b(z) = exp(-g_b(z)). \end{cases}$$

From this, we get he explicit solution $(V.2)$.

**Remark**: When $p=q=\frac{1}{2}$, this reduces to: $H^k(z)=f_b(z2^{-k})^{2^k}$. This case was extensively studied in [5, 17].

Now, we have:

$$H_n(u) = \sum_{k\geq 0} Proba(height=k/n\ keys)u^k = \sum_{k\geq 1}(H_n^k - H_n^{k-1})u^k + H_n^0$$

$$= (1-u)\sum_{k\geq 0} H_n^k u^k = (u-1)\sum_{k\geq 0}(1-H_n^k)u^k + 1$$

and (V.1) follows.

We claim[17] :

**Lemma V.2.** : *In any compact neighbourhood of* 0, *one has:* $g_b(z) = \frac{z^{b+1}}{(b+1)!} + O(z^{b+2})$.

Thus:

**Corollary**: $H^k(z) \sim exp\left(-\frac{z^{b+1}(p^{b+1}+q^{b+1})^k}{(b+1)!}\right)$ *when* $k \to \infty$.

As $H^k(z)$ is analytical, and $(1-H^k(z))u^k) \sim z^{b+1}\frac{(u(p^{b+1}+q^{b+1}))^k}{(b+1)!}$, we get the domain of analyticity of $H(z,u)$.

**Definition**: *We note:*

$$\begin{cases} F^k(z) = exp\left(-\frac{z^{b+1}(p^{b+1}+q^{b+1})^k}{(b+1)!}\right) \\ F(z,u) = (1-u)\sum_{k\geq 0} F^k(z)u^k = (u-1)\sum_{k\geq 0}(1-F^k(z))u^k + 1 \end{cases}$$

*$F(z,u)$ is analytical in the same domain as $H(z,u)$.*

## V.2. SOME ASYMPTOTICS:

We deal here with the asymptotic expansion of $H(z,u)$ when $u$ ranges in a compact neighbourhood of 1 included in $B(0, \frac{1}{p^{b+1}+q^{b+1}})$ and $z$ in $C_{0,\vartheta}$ (see Section 1). Here, $\vartheta$ cannot be arbitrary chosen as it will be of prime necessity to have $|f_b(z)| \leq 1$ or $Re(g_b(z)) \geq 0$. Such a restriction is minor, as it does not prevent from applying the tauberian Theorem to get $H_n(u)$ from $H(z,u)$ (see V.3. below).

We state our basic theorem of approximation:

**Theorem V.3.**: *There exists some constant c such that:*

$$H(z,u) = F(z,u) + O(|z|^{b+2}F(c|z|, |u(p^{b+2}+q^{b+2})|)),$$

*when $z$ ranges in $C_{0,\vartheta}$ and $u$ in some compact neighbourhood of* 1.

The proof is supported by two lemmas.

**Lemma V.4.**: *There exist three positive constants $A, B$ and $C$, and a cone $C_{0,\vartheta}$ such that:*

$$\begin{cases} Re(g_b(z)) \geq A|z|^{b+1} & ,|z| \leq 1 \\ Re(g_b(z)) \geq B|z| + C & ,|z| > 1. \end{cases}$$

**Lemma V.5.**: *There exist two positive constants $D$ and $E$ such that:*

$$H^k(z) = O(e^{-D|z|^{b+1}(p^{b+1}+q^{b+1})^k} + e^{-E|z|^{1-\alpha}})$$

*with:* $\alpha = \frac{1}{\log\frac{1}{q}} . \log\frac{q^b}{p^{b+1}+q^{b+1}}$, $0<\alpha<1$.

**Proof**: Let us note $M_z = E(\frac{\log z}{\log \frac{1}{q}})$. For $k > M_z$ one has: $|zp^{k_1}q^{k_2}| < |zq^k| < 1$ and using the first majoration in Lemma 1, one gets:

$$|H^k(z)| < e^{-A(p^{b+1}+q^{b+1})^k)|z|^{b+1}}>$$

For $k_1 = M_z$, let us define $K$ in $Z$ such that: $|zp^{k_1}q^{k-k_1}| \le 1$ iff $k_1 \ge K$. We also note: $S_1 = \sum_{k_1=0}^{K-1}\binom{k}{k_1}p^{k_1}q^{k-k_1}$ and $S_2 = \sum_{k_1=0}^{K-1}\binom{k}{k_1}p^{k_1}q^{k-k_1 b+1}$ Then, applying Lemma V.4.:

$$\log|H^k(z)| < -A|z|^{b+1}S_1 - (B|z|+c)S_2 .$$

Now, at least one of the following equations holds:

$$S_1 \ge |z|^{-\alpha} \tag{V.3}$$

$$S_2 \ge (p^{b+1}+q^{b+1})^k)(1-\beta) = (p^{b+1}+q^{b+1})^k)(1-\frac{p^{b+1}+q^{b+1}}{q^b}). \tag{V.4}$$

If ($V$.3) is not satisfied, one has indeed:

$$S_2 \ge (p^{b+1}+q^{b+1})^k) - max(p^{k_1}q^{k-k_1})S_1 \ge (p^{b+1}+q^{b+1})^k)(1-\left(\frac{q^b}{p^{b+1}+q^{b+1}}\right)^k z^{-\alpha})$$
$$\ge 1-\beta.$$

**Proof of Theorem V.3.**: We write:

$$H(z,u)-F(z,u) = H^{<1>}(z,u) + H^{<2>}(z,u) ,$$

with:

$$\begin{cases} H^{<1>}(z,u) = \sum\limits_{|z|^{b+2}(p^{b+2}+q^{b+2})>1} [H^k(z)-F^k(z)]u^k, \\ H^{<2>}(z,u) = \sum\limits_{|z|^{b+2}(p^{b+2}+q^{b+2})\le 1} [H^k(z)-F^k(z)]u^k, \end{cases}$$

First, we apply Lemma V.5. and write:

$$H^{<1>}(z,u) = O(\sum_{|z|^{b+2}(p^{b+2}+q^{b+2})>1} u^k(e^{-D|z|^{b+1}(p^{b+1}+q^{b+1})^k} + e^{-E|z|^{1-\alpha}})).$$

When $k$ ranges in this domain, *i.e.* $k < K_z = -\frac{(b+2)\log z}{\log(p^{b+2}+q^{b+2})}$, we have:

$$|z^{b+1}(p^{b+1}+q^{b+1})^k| > |z|^\gamma \text{ with } \gamma = b+2-(b+1)\frac{\log(p^{b+1}+q^{b+1})}{\log(p^{b+2}+q^{b+2})} > 0.$$

Moreover: $u^k < e^{K_z . max(\ln|u|)} = |z|^\delta$ with: $dd = \frac{(b+2)\max(\ln|u|)}{\log(p^{b+2}+q^{b+2})}$. Thus, setting $\varepsilon = inf(\gamma, 1-\alpha)$:

$$H^{<1>}(z,u) = O(|z|^\delta e^{-|z|^\varepsilon}).$$

Second, we consider $H^{<2>}(z,u)$. One sees easily that, for $k$ in this range, $|zp^{k_1}q^{k-k_1}| < 1$ and Lemma V.2. applies to $g_b(zp^{k_1}q^{k-k_1})$. Summing over $k$, one gets:

$$H^k(z) = e^{-\frac{z^{b+1}(p^{b+1}+q^{b+1})^k}{(b+1)!} + O(|z|^{b+2}(p^{b+2}+q^{b+2}))}$$

$$H^k(z)-F^k(z) = e^{-\frac{z^{b+1}(p^{b+1}+q^{b+1})^k}{(b+1)!}}.O(|z|^{b+2}(p^{b+2}+q^{b+2})).$$

As $Re(z^{b+1})>c\,|z|^{b+1}$, we get, after summation:

$$H^{<2>}(z,u) = O(|z|^{b+2}F(c\,|z|,|u(p^{b+2}+q^{b+2})|)).$$

■

We study now asymptotics of $F(z,u)$, when $z\to\infty$ inside $C_{0,\vartheta}$. We make use of *complex* Mellin transform:

$$F^*(s,u) = \int_0^\infty z^{s-1}(F(z,u)-1)dz\ ,$$

where the integration is done along any half line from 0 in $C_{0,\vartheta}$.

**Proposition V.6.:** *The Mellin transform $F^*$ is defined and analytical for: $-1<Re(s)<inf\,(0,\frac{\log|u|}{\log p^{b+1}+q^{b+1}})$ with:*

$$F^*(s,u) = [(b+1)!]^{\frac{s}{b+1}}\frac{(1-u)\Gamma(\frac{s}{b+1})}{1-u(p^{b+1}+q^{b+1})^{-\frac{s}{b+1}}}$$

**Proof:** We can develop formally:

$$F^*(s,u) = (u-1)\sum_k u^k(b+1)!^{\frac{s}{b+1}}(p^{b+1}+q^{b+1})^{-\frac{ks}{b+1}}\int_0^\infty z^{s-1}(1-e^{-z^{b+1}})dz$$

$$= \frac{(b+1)!^{\frac{s}{b+1}}}{b+1}.(u-1).\sum_k[u(p^{b+1}+q^{b+1})^{-\frac{s}{b+1}}]^k.(-\Gamma(\frac{s}{b+1}))\ ,$$

provided that both integral and Dirichlet geometric series are ***absolutely*** convergent.

**Proposition V.7.:** *Let $z$ and $u$ range respectively in $C_{0,\vartheta}$ and in any compact neighbourhood of 1 included in $B(0,(p^{b+1}+q^{b+1})^{-1})$. Then $F(z,u)$ satisfies asymptotically:*

$$F(z,u) = z^{-\frac{t(b+1)}{\log(p^{b+1}+q^{b+1})}}G(\log z,t) + O(z^{-m})$$

*where the periodic function $G$ is defined by:*

$$G(x,t) = \frac{(u-1)}{\log(p^{b+1}+q^{b+1})}\sum_{k\in Z}\Gamma\left(\frac{t+2ik\pi}{\log(pqb)}\right)e^{-\frac{2ik\pi}{\log(p^{b+1}+q^{b+1})}((b+1)x-\log(b+1)!)}$$

$t = \ln u$ ,

*$O(z^{-m})$ is uniform in the cone. This expression is to be continued when $u=1$ and in the paradoxal case $u=0$.*

It follows:

**Theorem V.8.:** $H(z,u) = z^{-\frac{t(b+1)}{\log(p^{b+1}+q^{b+1})}}(G(\log z,t)+O(z^{-\gamma})$ *with* $\gamma = (b+2)-(b+1)\frac{\log(p^{b+2}+q^{b+2})}{\log(p^{b+1}+q^{b+1})}>0$ , *$G$ defined as above, $z$ in $C_{0,\vartheta}$ and $u$ in a* *neighbourhood of* 1 *not containing* 0.

**Interpretation**: The distributions defined by $F$ and $H$, shifted by $\frac{(b+1)\log z}{\log(p^{b+1}+q^{b+1})}$, is asymptotically the one associated to the characteristic function $G(\log z, t)$. $G$ is a periodic function of $\log z$ with period $\frac{\log(p^{b+1}+q^{b+1})^{-1}}{b+1}$. Thus, $F$ and $H$ have for mean:

$$-\frac{(b+1)\log z}{\log(p^{b+1}+q^{b+1})} + P(\log z) + O(z^{-m})$$

and for variance: $Q(\log z) + O(z^{-m})$, when $P$ and $Q$ are periodic with period $\frac{\log(p^{b+1}+q^{b+1})}{b+1}$.

**Proof of Proposition V.7.**: This is steadily proved [7, 8]. All poles are simple when $u \neq 1$. These poles are 0 (with residue 1) and $\{\frac{-\log u+2ik\pi}{\log(p^{b+1}+q^{b+1})}\}_{k\in Z}$.

**Proof of Theorem V.8.**: From Theorem V.3. , one has:

$$H(z,u) = F(z,u) + O(|z|^{b+2}F(c|z|,|u(p^{b+2}+q^{b+2})|)).$$

As $G$ and $\ln u$ are bounded, we get from Proposition V.7.:

$$|z|^{b+2}F(c|z|,|u(p^{b+2}+q^{b+2})|) = O(|z|^{-\frac{(b+1)(\ln u+\log(p^{b+2}+q^{b+2}))}{\log(p^{b+1}+q^{b+1})}+b+2})+O(z^{b+2-m})$$

$$+ O(|z^{-\frac{(b+1)\ln u}{\log(p^{b+1}+q^{b+1})}}|.|z|^{-\gamma}).$$

### V.3. BACK TO BERNOULLI CASE:

We apply here the semi tauberian theorem developped in the third section.

**Theorem V.9.**: *The distribution of the height of tries with $n$ keys converges* in the sense of distributions *to the distribution with characteristic function* $G(n,u)$.

**Proof**: One would like to prove directly:

$$H_n(u)-G(n,u) \underset{n\to\infty}{\to} 0.$$

Unfortunately, it appears difficult to establish the first regularity condition for $H_n(e^{it})$, with $t \in \mathbb{R}$. Still, it is possible to extract a decreasing real positive sequence and to rely on the second regularity condition. Indeed, the sequence $\{H_n^k\}_{n\in N}$ is decreasing (a new insertion can only increase the height). Thus, one may readily claim:

$$H_n^k - exp\,(-(p^{b+1}+q^{b+1})^k)\frac{n^{b+1}}{(b+1)!}) \underset{n\to\infty}{\to} 0$$

for functions $exp\,(-(p^{b+1}+q^{b+1})^k)\frac{x^{b+1}}{(b+1)!})$ clearly have a normal variation. In the same manner, let us note $\Delta_n$ and $\Delta(z)$ the discrete distributions:

$$\begin{cases} \Delta_n = \sum_{k\in N} H_n^k \delta_k \\ \Delta(z) = \sum_{k\in N} H^k(z)\delta_k \end{cases}$$

Let $\varphi$ be any test positive function with compact support such that $\int \varphi = 1$. It is also clear that the sequence $\{<\Delta_n, \varphi>\}_{n \in N}$ is positive, decreasing and upper bounded by 1 and that $<\Delta(x), \varphi>$ is a normally varying function. Thus:

$$<\Delta_n \,|\, \varphi> - <\Delta(n) \,|\, \varphi> \underset{n \in N}{\rightarrow} 0 \,,$$

whatever the test function $\varphi$ is. And the theorem is proved.

**Remark**: The size of a trie is also a parameter with an increasing sequence $\{P_n^k\}$. We could also use this reasoning, and the semi-tauberian theorem, to prove the convergence to the normal law of this parameter in the Bernoulli case.

**VI. CONCLUSION:**

Studying Limiting Distributions provides an accurate and systematic method for asymptotical analysis of algorithmical parameters. The systematical use of generating functions and the reference to complex analysis prove to be a valuable tool. We derive a semi-tauberian theorem which is a powerful tool to translate results obtained under the approximate Poisson model to the Bernoulli model. Thus, its field of applications is larger than the one presented here. We also exhibit limiting distributions. In particular, the asymptotic values for the moments of any order are steadily derived from these limiting distribution. This generalizes and completes previous results as we get all asymptotics for means and variances for both Poisson and Bernoulli models, in the uniform and biased case. Moreover, we discovered new phenomena that do not appear when the analysis is restricted to the average values. For example, when dealing with the depth of insertion, we point out a solution of continuity between the biased and uniform case, although the mean is $O(\log n)$ in both cases. One finds a gaussian dilatation around the mean in the biased case while in the uniform case the distribution remains centered around its mean with a $O(1)$ deviation.

Researchs in trie parameters area are not achieved. We are actually working at the following topics.

* We study the distribution of the depth of insertion for a given key (a fixed sequence of bit) when inserted among $n$ other random keys. When $p \neq q$, the gaussian behaviour mentioned above disappears and gives place to a centered distribution around a mean $O(\log n)$ determined by the sequence of bits of the pointed key.
* We extend the previous results to some *markovian* process. Markovian process can be a good modellization when the keys are alphabetical. The transition matrix is then formed with the transition probabilities from one letter to another. Our analysis is a particular case of a stationary process over a binary alphabet.

**References**

1. Ph. Flajolet and C. Puech, "Tree Structure for Partial Match Retrieval," pp. 282-288 in *Proc. 24-th I.E.E.E. Symp. on FOCS*, (1983). INRIA Reseach Report,233,submitted

2. R. Fagin, J. Nievergelt, N. Pippenger, and H.R. Strong, "Extendible Hashing:A Fast Access Method for Dynamic Files," *ACM TODS* **4,3** pp. 315-344 (1979).

3. G. Fayolle, Ph. Flajolet, M. Hofri, and Ph. Jacquet, "Analysis of a Stack Algorithm for Random Multiple-Access Communication," *IEEE Trans. on Information Theory* **IT-31,2** pp. 244-254 (1985).

4. W. Feller, *An Introduction to Probability Theory and its Applications*, Wiley-third Edition-1971 (1957).

5. Ph. Flajolet, "On the Performance Evaluation of Extendible Hashing and Trie Searching," *Acta Informatica* **20** pp. 345-369 (1983).

6. Ph. Flajolet, M. Régnier, and D. Sotteau, "Algebraic Methods for Trie Statistics," *Annals of Discrete Mathematics* **25** pp. 145-188 (1985).

7. Ph. Flajolet, M. Regnier, and R. Sedgewick, *Some Uses of the Mellin Transform Techniques in the Analysis of Algorithms* , Springer NATO ASI SEr. F12, Combinatorial Algorithms on Words (1985).

8. Ph. Flajolet, M. Regnier, and R. Sedgewick, "Mellin Transform Techniques for the Analysis of Algorithms ," *Monography in preparation*, (1986).

9. Ph. Flajolet and N. Saheb, "Digital Search Trees and the Complexity of Generating an Exponentially Distributed Variate," in *Proc. Coll. on Trees in Algebra and Programming*, Lecture Notes in Computer Science, L'Aquila (1983). to appear

10. Ph. Flajolet and J.M. Steyaert, "A Branching Process Arising in Dynamic Hashing, Trie Searching and Polynomial Factorization," pp. 239-251 in *Proceedings ICALP 82* , Lecture Notes in Computer Science (1982).

11. D. Knuth, *The Art of Computer Programming*, Addison-Wesley,Reading,Mass. (1973).

12. D. LAZARD, *On Polynomial Factorization*. 1982.

13. P.A. Larson, "Dynamic Hashing," *BIT* **18** pp. 184-201 (1978).

14. W. Litwin, "Trie Hashing," pp. 19-29 in *Proc. ACM-SIGMOD Conf. on MOD,*, , Ann Arbor,Mich. (1981).

15. J. Nievergelt, H. Hinterberger, and K.C. Sevcik, "The Grid-file: an Adaptable Symmetric Multi-Key File Structure," *ACM TODS* **9,1**(1984).

16. N.E. Norlund, *Vorlesungen Uber Differenzenrechnung*, Chelsea Publishing Company (1954).

17. M. Régnier, "Evaluation des performances du hachage dynamique," *These de 3-eme cycle,Universite d'Orsay*, (1983).

18. W. A. Rosenkrantz, *Collision Resolution Interval*, Univ. Massachusetts, Internal Report 1985.

# APPENDIX

## A- THE NEGLIGIBLE REMAINING TERMS THEOREM

**Theorem (A-1):** *Let $t$ be member of the support of $\varphi_n$ (of the form $[n-\beta\sqrt{n}, n+\beta\sqrt{n}]$, $\beta$ being a positive real), let $m$ be such that $|m-n|>\alpha\sqrt{n}\log n$, then for every integer $k$:*

$$Z[f_m](t) = O((m-t)^{-k})$$

*when $n \to \infty$.*

**Proof** (beginning):

According to (2-3) we have the equation:

$$Z[f_m](t) = \frac{1}{2i\pi}\oint \frac{1}{(1+i\omega)^{m+1}} e^{t(i\omega+\omega^2/2)} d\omega ,$$

where the contour of integration is any loop $\Gamma$ containing the singularities of $f_m^{*}(\omega)$, in this particular case the singularities are restricted to $\{i\}$. For convenience of writing we will study $Z[f_{m-1}](t)$ in order to avoid $m+1$ in the right-hand expressions.

**Lemma (A-2):** *we can choose $\Gamma$ in the set*

$$I = \{\omega/ \ |1+i\omega|>1\}\cup\{\omega/ \ |\exp(i\omega+\frac{\omega^2}{2})|<1\}\cup\{0\} .$$

**Proof:**

$\{\omega / \ |1+i\omega|>1\}$ is the complementary of the unit disk centered on $i$. $\{\omega / \ |\exp(i\omega+\frac{\omega^2}{2})|<1\}$ is equal to $\{\omega / \ \mathrm{Re}(i\omega+\omega^2/2)<0\}$ which corresponds to the two parts of the complex plan above and below the hyperbol $\{i+\omega=x+iy/ \ x^2-y^2+1=0\}$. The unit disk centered on $i$ is included in the above part of the hyperbol. In fact if $x^2-y^2+1=0$ then

$$x^2+(y-2)^2 = y^2-1+(y-2)^2 > 1$$

except the case of equality when $\omega=0$. ■

**Lemma (A-3):** The parametrisation of $\Gamma$. *Let $\tau$ be a real parametrisation of $\Gamma$. There exists $\Gamma$ in $I$ and a neighbourhood of 0, $[a,b]$, such that*

$$\tau\in[a,b] , \ 1+i\omega=e^{i\vartheta} , \ \vartheta=\tau-\frac{\tau^3}{6}-\frac{17}{48}i\tau^4 ,$$

*and* $0<D<1$

$$Z[f_{m-1}](t) = \frac{1}{2i\pi}\int_a^b e^{mi\vartheta}\exp\{t(e^{i\vartheta}-1)-t\frac{(e^{i\vartheta}-1)^2}{2}\}\frac{d\omega}{d\tau}d\tau+O(D^{m+t}) .$$

**Proof:**

taking $\vartheta=\tau-\frac{\tau^3}{6}-\frac{17}{48}i\tau^4$ we get the expansion near 0 of the analytical functions:

$$(1+i\omega)^{-1} = \exp\{-i\tau+i\frac{\tau^3}{6}-\frac{17}{48}\tau^4\}$$

$$\exp\{i\omega+\frac{\omega^2}{2}\} = \exp\{-i\tau+i\frac{\tau^3}{6}-\frac{17}{48}\tau^4+l(\tau)\} \tag{4-3}$$

with $l(\tau) = O(\tau^5)$. So there is an interval $[a,b]$ containing 0 where $\omega = \omega(\tau)$ is element of $I$. Outside $[a,b]$ we complete $\Gamma$ in such a way that there exists $0<D<1$ such that $|1+i\omega|$ and $|\exp(i\omega+\omega^2/2)|$ are not greater than $D$ (this is always possible because $I-\{0\}$ is an open area.

**Proof of of the theorem** (continuing)

Inside $[a,b]$ we can write, with the homothetic change of variable $(m+t)\tau^4 = \tau'^4$:

$$\int_a^b (1+i\omega)^{-m} e^{t(i\omega+\omega^2/2)} \frac{d\omega}{d\tau} d\tau =$$

$$(m+t)^{-1/4} \int_{a(m+t)^{1/4}}^{b(m+t)^{1/4}} \exp\{\frac{(t-m)}{(t+m)^{1/4}} i\tau' + (m+t)^{1/4} \frac{i\tau'^3}{6}\} g_{m,t}(\tau') d\tau'$$

with

$$g_{m,t}(\tau') = \exp\{-\frac{17}{48}\tau'^4 + t\; l(\frac{\tau'}{(m+t)^{1/4}})\} \frac{d\omega}{d\tau}(\tau = \frac{\tau'}{(m+t)^{1/4}}) .$$

■

**Lemma (A-4)**: the $C^k$ prolongation of $g_{m,t}$, k fixed. *We can continue $g_{m,t}$ in $]-\infty, +\infty[$ with the following properties:*

(i) $g_{m,t}$ *is* $C^k(]-\infty, +\infty[)$.

(ii)

$$\int_{J_{m,t}} \exp\{\frac{(t-m)}{(t+m)^{1/4}} i\tau' + (m+t)^{1/4} \frac{i\tau'^3}{6}\} g_{m,t}(\tau') = O(E^{m+t})$$

*with* $J_{m,t} = ]-\infty, a(m+t)^{1/4}[\cup]b(m+t)^{1/4}, +\infty[$ *and* $0<E<1$.

(iii) local convergence: *on every compact set of $R$ and for every integer $l \le k$ we have the uniform convergence $g_{m,t}^{(l)} \to g^{(l)}$ when $m+t$ increases, with* $g(\tau) = \exp\{-\frac{17}{48}\tau^4\}$.

(iv) dominated convergence: *there exist a real $K>0$ such that, for every integer $l \le k$ and for all $\tau \in R$:*

$$|g_{m,t}(\tau)| \le K\, q_5(|\tau|) g(\tau)$$

*where $q_5$ is a positive coefficient polynomial of degree five.*

At first we introduce a convenient notation: $g(x) = O_k(f(x))$, $x \in J \subset R$, g being a real $C^k$ function and $f$ a $C^k$ positive function (*i.e.* for $x$ positive and for all integer $l \le k$: $f^{(l)}(x) \ge 0$, for example, $f$ can be a positive coefficient polynomial). This notation will mean:

$$\forall x \in J \,,\ \forall l \le k \ :\ g^{(l)}(x) = O(f^{(l)}(|x|)) \ \text{ or}$$

$$\exists K>0 : \ \forall x \in J \,,\ \forall l \le k \ : |g^{(l)}(x)| \le K f^{(l)}(|x|)$$

with the restricting and conventional statement:

$$\text{if } f^{(l)} = 0 \ \text{ then } \ g^{(l)}(x) = O(1) \,.$$

**Proof**

Because of the analyticity of all the above functions, $g_{m,t}$ and others, we can derive the Taylor expansions. Thus, for every k, according to (4-2), we can write:

$$\exp(i\omega + \frac{\omega^2}{2}) = \exp\{i\tau + i\frac{\tau^3}{6} - \frac{17}{48}\tau^4 + O_k(\tau^5)\}$$

$$\frac{d\omega}{d\tau} = 1 + O_k(\tau) \qquad \text{for } \tau \in [a,b] \,.$$

So we can write:

$$\int_{a(m+t)^{1/4}}^{b(m+t)^{1/4}} \exp\{\frac{(t-m)}{(t+m)^{1/4}} i\tau + (m+t)^{1/4}\frac{i\tau^3}{6}\} g_{m,t}(\tau) d\tau =$$

$$\int_{a(m+t)^{1/4}}^{b(m+t)^{1/4}} \exp\{\frac{(t-m)}{(t+m)^{1/4}} i\tau + (m+t)^{1/4}\frac{i\tau^3}{6} - \frac{17}{48}\tau^4 + \frac{tO_k(\tau^5)}{(m+t)^{5/4}} + \frac{O_k(\tau)}{(m+t)^{1/4}}\} d\tau$$

or more concisely

$$g_{m,t}(\tau) = \exp\{-\frac{17}{48}\tau^4 + \frac{1}{(m+t)^{1/4}} O_k(\tau+\tau^5)\} \,.$$

If we write $g_{m,t}(\tau) = \exp(-17\tau^4/48)(1+(m+t)^{-1/4} l_{m,t}(\tau))$ with always $l_{m,t}(\tau) = O_k(\tau+\tau^5)$, we can obviously continue $l_{m,t}$ to $J_{m,t}$, such that $l_{m,t}$ now for all $\tau$ real. For example, at raccording point $a(m+t)^{1/4} = ar$:

$$l_{m,t}(\tau) = l_{m,t}^{(0)}(ar) + \cdots + \frac{(\tau - ar)^5}{5!} l_{m,t}^{(5)}(ar) + O_k(1) \,,$$

and we can $C^k$-continue $O_k(1)$ with a linear combination of $k$ decreasing exponential with different rate.

This continuation done, the proof of the lemma becomes trivial. ■

**Notation:**

$$\eta = \left[\frac{|t-m|}{(m+t)^{1/2}}\right]^{1/2} ,$$

$$v = \operatorname{sign}(t-m) \,,$$

$$\xi = v\left[\frac{|t-m|^3}{m+t}\right]^{1/2} .$$

**Lemma (A-5):** *$\eta$ and $\xi$ growth to infinity as $|m-n|/\sqrt{n}$ increases. we also have the following minorations:*

$$\eta^2 > |t-m|^{1/2} \cdot \frac{1}{\sqrt{n}} \left[\frac{1}{2}\min\{\sqrt{n}\,,\ \frac{|m-t|}{\sqrt{n}} \cdot \frac{n}{2t}\}\right]^{1/2}$$

$$\xi^2 > \sqrt{n}\left(\frac{m-t}{\sqrt{n}}\right)^2 . \frac{1}{2}\min\{\sqrt{n}\,,\ \frac{|m-t|}{\sqrt{n}}.\frac{n}{2t}\}\,.$$

**Proof**: We will compute minoration of $\eta$ and $\xi$.

$$\eta^2 = \frac{|t-m|}{\sqrt{n}}\left(\frac{n}{m+t}\right)^{1/2} = \left(\frac{|t-m|}{\sqrt{n}}\right)^{1/2}\left(\frac{1}{\sqrt{n}} + \frac{2t}{n}.\frac{\sqrt{n}}{m-t}\right)^{-1/2}$$

$$\xi^2 = \left(\frac{|t-m|}{\sqrt{n}}\right)^3\sqrt{n}\left(\frac{n}{m+t}\right) = \left(\frac{t-m}{\sqrt{n}}\right)^2\sqrt{n}\left(\frac{1}{\sqrt{n}} + \frac{2t}{n}.\frac{\sqrt{n}}{m-t}\right)^{-1}.$$

We conclude with the general inequality:

$$(a,b)\in R\times R: \quad \frac{1}{|a+b|} > \frac{1}{2}\min\{\frac{1}{|a|},\ \frac{1}{|b|}\}\,.$$

**Remark**: we have $\frac{t}{n} = 1 + O(\frac{1}{\sqrt{n}})$ then asymptotically:

$$\eta^2 > |t-m|^{1/2}.\frac{1}{\sqrt{n}}\left(\frac{1}{2}\min\{\sqrt{n}\,,\ \frac{|m-t|}{\sqrt{n}}\}\right)^{1/2}$$

$$\xi^2 > \sqrt{n}\left(\frac{m-t}{\sqrt{n}}\right)^2.\frac{1}{2}\min\{\sqrt{n}\,,\ \frac{|m-t|}{\sqrt{n}}\}\,. \qquad \blacksquare$$

**Proof of the theorem** (continuing)

We obtain a new expression with the homothetic change of variable $\eta\tau' = \tau''$:

$$Z[f_{m-1}](t) = \frac{v}{2i\pi}\left(\frac{|t-m|}{m+t}\right)^{1/2}\int_{-\infty}^{+\infty}\exp\{i\xi(\tau'' + v\frac{\tau''^3}{6})\}g_{m,t}(\eta\tau'')d\tau''$$

$$+ O(E^{m+t})\,.$$

Thus it is obvious that we have to study general integrations such

$$\int_{-\infty}^{+\infty}\exp\{-i\xi(x+v\frac{x^3}{6})\}g(\eta x)dx$$

where $g$ is a $C^k$ fast decreasing function (i.e $\forall\ l\le k,\ \forall\ x^{k'}g^{(l)}(x)\to 0$ when $|x|\to\infty$). We just mention the following lemma.

**Lemma (A-6)**: Theorem of the stationary phase. *Let $g$ be a $C^k(R)$ fast decreasing function, $P$ a* real *polynomial, its derivative $P'$ having $x_1,\cdots,x_n$ different roots of respective multiplicity $d_1,\cdots,d_n$ (thus the degree of $P$ is $d_1+\cdots+d_n+1$).*

$$G(\xi) = \int_{-\infty}^{+\infty}e^{-i\xi P(x)}g(x)dx$$

*has the asymptotic expansion:*

$$G(\xi) = G_1(\xi)+\cdots+G_n(\xi)+G_0(\xi)\,.$$

*where*

$$|G_0(0)| \le A_{0,l} \cdot \frac{1}{\xi^l} \|g\|_l ;$$

*and, for $i \ge 1$, $G_i$ is the contribution of the $i^{th}$ root $x_i$:*

$$G_i(\xi) = \begin{array}{ccc} \dfrac{B_{i,1,0}(g)}{\xi^{\frac{1}{d_i+1}}} + \cdots + \dfrac{B_{i,d_i,0}(g)}{\xi^{\frac{d_i}{d_i+1}}} \\ + \quad \cdots \quad + \\ \dfrac{B_{i,1,l-1}(g)}{\xi^{l-1+\frac{1}{d_i+1}}} + \cdots + \dfrac{B_{i,d_i,l-1}(g)}{\xi^{l-1+\frac{d_i}{d_i+1}}} + O(\dfrac{1}{\xi^l}) \end{array}$$

*with*

$$|O(\frac{1}{\xi^l})| \le A_{i,l} \cdot \frac{1}{\xi^l} \|g\|_{(d_i+1)l} .$$

*$B_{i,d',l'}(g)$ are linear combinations of the $d'+l(d_i+1)$ first derivations of $g$ at point $x_i$ and the coefficients $A_{i,l}$ are independent from $g$. We note:*

$$\|g\|_l = \|g\|_0 + \cdots + \|g^{(l)}\|_0 .$$

*with* $\|g\|_0 = \int_{-\infty}^{+\infty} |g|$.

If we take $P(x) = x + \frac{x^3}{6}$, then $P(x)$ has no real root and we can evaluate

$$\int e^{-i\xi(x+x^3/6)} g(\eta x) dx = G_0(\xi) ,$$

with

$$|G_0(\xi)| \le A_{0,l} \left(\frac{\eta}{\xi}\right)^l \|g\|_l$$

as $\eta$ and $\xi$ increase to infinity (more rigously, $\xi$ greater outside of an arbitrary neighbourhood of 0), l being an arbitrary integer. Then we can state the lemma:

**Lemma (A-7)**: *if $m-n < -\alpha\sqrt{n}\log n$ we have*

$$Z[f_m](t) = O\left(\left(\frac{\sqrt{n}}{t-m}\right)^{\frac{k}{2}-1} (t-m)^{\frac{k+1}{2}}\right)$$

**Proof**: we can deduce from the analysis done above

$$Z[f_{m-1}](t) = \frac{v}{2i\pi}\left(\frac{|t-m|}{m+t}\right)^{1/2} \int_{-\infty}^{+\infty} e^{i\xi(x+x^3/6)} g_{m,t}(\eta x) dx + O(E^{m+t})$$

$$= O\left(A_{0,l} \cdot \left(\frac{|t-m|}{m+t}\right)^{1/2} \left(\frac{\eta}{\xi}\right)^k \|g_{m,t}\|_k\right)$$

$$= O\left(\left(\frac{|t-m|}{m+t}\right)^{1/2}\left(\frac{(m+t)^{1/2}}{(t-m)^2}\right)^{k/2}\right)$$

because $\| g_{m,t} \|_k \| = O(1)$

$$= O\left(\left(\frac{m+t}{n}\right)^{\frac{k}{4}-\frac{1}{2}} \cdot \left(\frac{\sqrt{n}}{t-m}\right)^{\frac{k}{2}-1} \cdot (t-m)^{\frac{k+1}{2}}\right).$$

With $\frac{m+t}{n} = O(1)$ we achieve the proof.

If we take $P(x) = x - \frac{x^3}{6}$, $P(x)$ has two real roots, $-\sqrt{2}$ and $\sqrt{2}$, then

$$\int e^{-i\xi(x-x^3/6)} g(\eta x)dx = G_1(\xi) + G_2(\xi) + G_0(\xi)$$

with always $|G_0(\xi)| \le A_{0,l}\left(\frac{\eta}{\xi}\right)^l \|g\|_l$, and $G_1(\xi)$ and $G_2(\xi)$ as linear combination of the $2l$ derivatives of $g(\eta x)$ at $x = \pm\sqrt{2}$ combined with powers of $\xi^{-1/2}$, plus an approximation of $O(\eta^{2l}/\xi^l \|g\|_{2l})$. So we can state the lemma

**Lemma (A-7)**: *if* $m-n > \alpha\sqrt{n}\log n$ *we have*

$$Z[f_m](t) = O\left((m+t)^{-1/2}.|t-m|^{\frac{k-1}{2}}\right)$$
$$+O\left(\left(\frac{|t-m|}{\sqrt{n}}\right)^{\frac{3}{2}(2k+5)} \cdot \left(e^{-\frac{17}{48}(m-t)} + e^{-\frac{17}{48}\left(\frac{m-t}{\sqrt{n}}\right)^2}\right)\right).$$

**Proof**: we have to proove this asymptotic only for $G_1$ and $G_2$, according to the above notations, because $G_0$ is already analysed in the previous lemma. Thus we have the expression

$$Z[f_{m-1}](t) = \frac{v}{2i\pi}\left(\frac{|t-m|}{m+t}\right)^{1/2}\int_{-\infty}^{+\infty} e^{i\xi(x-x^3/6)} g_{m,t}(\eta x)dx + O(E^{m+t})$$

$$= O(\eta^{2k+5} g_{m,t}(\pm\sqrt{2}\eta))$$

$$+O\left(\left(\frac{|t-m|}{m+t}\right)^{1/2}\left(\frac{\eta^2}{\xi}\right)^k \|g_{m,t}\|_{2k}\right)$$

$$+O\left(\left(\frac{|t-m|}{m+t}\right)^{1/2}\left(\frac{\eta}{\xi}\right)^k \|g_{m,t}\|_k\right).$$

The first right-hand term is a rough majoration of what could be a ***linear combination of the first 2l derivatives of*** $g(\pm\sqrt{2}\eta)$ ***mixed with powers of*** $\xi^{-1/2}$. The second one is the remaining terms of $G_1(\xi)$ and $G_2(\xi)$. The last one is of course $G_0(\xi)$.

We achieve the proof with $\frac{\eta^2}{\xi} = \frac{1}{|t-m|^{1/2}}$ and with the minorations of the lemma (A-5) which helps reaching the evaluation

$$g_{m,t}(\pm\sqrt{2}\eta) = O\left(\eta^5\left(e^{-\frac{17}{48}(m-t)} + e^{-\frac{17}{48}\left(\frac{m-t}{\sqrt{n}}\right)^2}\right)\right),$$

and the rough majoration $\eta \le \left(\frac{|m-t|}{\sqrt{n}}\right)^{1/2}$.

■

**Proof of the theorem** (end): now in order to achieve the proof of the main theorem we claim that we can tune the two previous lemma to, for an arbitrary k, the general evaluation:

$$Z[f_m](t) = O\left((m-t)^{-k}\right)$$

when $|m-n| > \alpha\sqrt{n}\,\log n$, as $n \to \infty$.

- It is obvious for the first one.
- For the second one we need to consider the behaviour of the exponentials. We have obviously

$$e^{-\frac{17}{48}(m-t)} = O\left((m-t)^{-k}\right).$$

and after a little refexion we have also

$$e^{-\frac{17}{48}\left(\frac{m-t}{\sqrt{n}}\right)^2} = O\left((m-t)^{-k}\right),$$

because asymptotically

$$\left(\frac{m-t}{\sqrt{n}}\right)^2 > \alpha\log n \,.\, \frac{m-t}{\sqrt{n}},$$

and

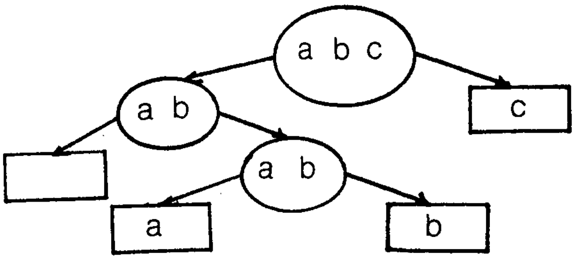
$$k\log(m-t) = k\left(\log\left(\frac{m-t}{\sqrt{n}}\right) + \frac{1}{2}\log n\right).$$

■

## B- NORMAL VARIATION OF THE DISTRIBUTION OF SIZE OF THE TREES

### B- 1 An example

Let's consider this example of trie




with the keys: a=101100.., b=10000... and c=0010...
Let's operate the insertion of a new key d.

* If d=11..., d occurs in the empty leaf then the size of the tree does not vary.
* If d=0...., d occurs in the leaf of c and causes its partitioning.

### B- 2 General case

Let $\mathbf{X}_n$ be the random variable of the size of the tree, conditioned by the number $n$ of key. We have the relation

$$P_n^k = \mathrm{E}[\delta(\mathbf{X}_n = k)] \, ,$$

where $\delta(true) = 1$ and $\delta(false) = 0$, and E denote the expectation of a random variable. $\delta(\mathbf{X}_n = k)$ is the characteristic variable associated to the event *size of the tree equals $k$* conditioned by $n$.

Let $\mathbf{r}_n^0$ be the characteristic variable associated to the event *the $n+1^{th}$ insertion occurs in a non saturated leaf* (the cardinality of the leaf is less than $b$). Let $\mathbf{r}_n^1$ be the characteristic variable associated to the opposite event. We have the relation

$$\mathbf{r}_n^1 + \mathbf{r}_n^0 = 1$$

Let's note $P_n^{0,k} = \mathrm{E}[\mathbf{r}_n^0 \delta(\mathbf{X}_n = k)]$ and $P_n^{1,k} = \mathrm{E}[\mathbf{r}_n^1 \delta(\mathbf{X}_n = k)]$. We have

$$P_n^{1,k} + P_n^{0,k} = P_n^k \, .$$

### B- 3 Application to the generating functions.

Thus if $P_n^0(u) = \sum_k P_n^{0,k} u^k$ and $P_n^1(u) = \sum_k P_n^{1,k} u^k$ we have

$$P_n^0(u) + P_n^1(u) = P_n(u) \, .$$

We know that an insertion in a non saturated leaf does not modify the size of the tree. An insertion in a saturated leaf causes its partitioning and increases the tree by adding at this leaf a new sub-tree with a root of cardinality $b+1$, and it takes place independently of the environing tree structure (it is independent of the *context* if we use the vocabulary of the Terms Algebra). Thus we can formaly write

$$X_n = r_{n_0}X_n + r_n^1 X_n$$
$$X_{n+1} = r_n^0 X_n + r_n^1 \, (X_n + (X_{b+1} - 1))$$

(the sign + are designed for different purposes, the first one divides opposite events the second separates independent events). This relation is translated for the generating functions into

$$P_n(u) = P_n^0(u) + P_n^1(u)$$
$$P_{n+1}(u) = P_n^0(u) + (P_{b+1}(u) . \frac{1}{u}) P_n^1(u) \, .$$

Thus we can evaluate

$$P_{n+1}(u) - P_n(u) = (1 - \frac{1}{u} P_{b+1}(u)) P_n^1(u) \, .$$

When $|u| \leq 1$ we have $|P_n^1(u)| \leq 1$. Thus, with $u = e^{-it}$, $t$ being real,

$$P_{n+1}((u) - P_n(u) = O(1 - e^{it} P_{b+1}(e^{-it})) = O(t)$$

uniformly with respect to $n$ when $t$ is element of a real compact neighbourhood of 0 included in the set of definition of $P_{b+1}(e^{-it})$. ■

Imprimé en France

par

l'Institut National de Recherche en Informatique et en Automatique